Multimodal Fusion Towards Crime Prevention on the Edge

Amna Anwar

N0718978

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

July, 2023

Abstract

Detecting violent language is a complex problem in preventing crime and harmful content. Violent language detection in real-time conversations is therefore a novel problem in computer science, with most current solutions focusing on the either text-based or audio-based solutions. These solutions will often miss the wider context, without audio it is difficult to extract auditory features, and without text it is difficult to understand the language used. In addition, there has been growing interest in the use of edge computing technologies to prevent crime. Edge computing is the processing of data at or close to the edge of the network, as opposed to sending it to a centralised data centre. Faster response times, lower bandwidth needs, and improved data security are just a few benefits of this strategy for preventing crime, which when combined with a multimodal dataset could achieve improved detection while preserving user privacy.

This thesis investigates the practical application of multimodal data fusion and edge computing for crime prevention, specifically focussing on the detection of violent language in conversations from text and audio data. A fusion algorithm that combines natural language processing (NLP) techniques of Bidirectional Encoder Representations from Transformers (BERT) and Linguistic Inquiry and Word Count (LWIC), in addition to Mel-frequency cepstral coefficients (MFCC) and time-frequency domain features was developed. The resulting F1 score of 0.85 demonstrates the effectiveness of the algorithm in identifying potential instances of violent conversations related to domestic violence or public safety when compared to single modality results. However, the initial iteration of the algorithm required substantial

computational resources, leading to its compression using model reduction for deployment on edge devices such as mobile phones and smart home devices.

To facilitate real-time detection, a mobile application and a cost-effective smart home device were developed, utilising a model reduction approach. The mobile application enables timely identification of violent conversations, while the smart home device serves as an alternative for people without access to mobile phones. The approach gives consideration to contextual factors such as microphone quality and device positioning, which influence the algorithm's adaptability to different scenarios. Future research aims to enhance the accuracy of the model, improve the realism of training data, and explore innovative approaches for contextual analysis and result normalisation. This thesis contributes to the advancement of multimodal technologies for crime prevention, highlighting the importance of data fusion and edge devices in this domain.

Acknowledgements

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ

I am honored to extend my heartfelt appreciation to my supervisor, Professor Eiman Kanjo, who has been a constant source of inspiration to me, starting from my first year of bachelor's, as a remarkable role model for women excelling in STEM fields. Her exceptional achievements as a female researcher of colour in STEM have fuelled my aspirations to strive for success in my own path. I am truly grateful for her mentorship, which not only has nurtured my growth as a researcher, but has also instilled in me the confidence and determination to make meaningful contributions in my chosen field.

I am sincerely grateful to Dr. Andreas Oikonomou for his support and guidance as my supervisor for both my Bachelor's final-year project and my PhD. His belief in my abilities and his encouragement to pursue a Ph.D. have played a pivotal role in my academic journey.

This journey has been more enjoyable thanks to the support and constant encouragement of the past and present members Smart Sensing Lab. I also thank Dr. Dario Anderez for the mentorship and guidance he provided at the beginning of my Ph.D. and also for the ongoing support that he provided after leaving the lab. I also express my sincere gratitude to Dr.(to be) James Williams for his steadfast support and guidance throughout my degree.

All my accomplishments are due to the tireless support and courage of my mother, Rizwana Anwar. Her decision to leave Germany and move to England, solely to create better opportunities for me, is a testament to her selflessness and love. She is undeniably my superhero and I will

forever be grateful for her boundless efforts and sacrifices. Equally, my father, Anwar-Ul-Haq, has consistently been my biggest source of encouragement. He has continuously pushed me to exceed my own expectations and believed in my abilities even when I doubted myself. I consider myself extremely fortunate to have parents like mine, whose selfless support and belief in me have shaped my path and fuelled my achievements. I am truly blessed to be their daughter. Also, not to forget my most favourite beings, my siblings, Saad and Maryam, and my emotional support cats Shahjahan and Bano for their constant love and support.

Finally, I would like to acknowledge and thank my extended family and close friends for their constant support and prayers. I appreciate you all, I would not have been able to get this far without their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview and Motivation

The World Health Organisation (WHO) found that nearly 1 in 3 (30%) women have been subjected to physical and / or sexual violence in a prevalence data survey between 2000-2018 in 161 countries and areas [4]. A report referenced by the WHO [5] also identified the health issues of domestic violence, with 42% of women reporting injury as a consequence of violence against women [6]. The WHO identifies several factors associated with violence against women, including lower levels of education, harmful use of alcohol, harmful masculine behaviours, and low levels of gender equality [5]. Violence against men has also been studied in the domestic setting, with prevalence rates of 3.4% to 20.3% for physical violence against men surveyed in this setting [7]. This highlights the issue of crime that still occurs at high rates around the world, despite the reported levels of crime in most countries decreasing on average [8]. In addition to current statistics on domestic violence worldwide, some authors [9] have considered how increased social isolation and the inability to access potential support resources have increased the risk of domestic violence becoming a public health crisis. For example, a meta-analysis identified how, in a systematic review of the literature, stay-at-home orders related to the Covid-19 pandemic increased domestic violence incidents [10]. Therefore, it is necessary to consider more localised statistics and to consider how research can positively improve the ability of individuals to report

and identify domestic violence.

In England and Wales, details of domestic abuse are published in a statistical bulletin from the Crime Survey for England and Wales (CSEW) by the Offline for National Statistics (ONS) [11]. The CSEW estimated that 5% of adults (6.9% women and 3% men) aged over the age of 16 experienced domestic abuse during the year ending March 2022 [11]. There was no change in the prevalence of cases of domestic abuse experienced by adults, compared to the 8% decrease in estimated crime for the year ending June 2022. Although the terminology differs from the previously identified WHO report on violence against women [4], the CSEW approximates that 1 in 5 adults aged 16 and over (10.4 million) had experienced domestic abuse as adults [11]. The CSEW does report that the number of domestic abuse police cases had increased by 7.7% from the previous year to a total of 910,980 reported incidents, following similar increases in previous years [11]. The Opinions and Lifestyle Survey (OPN) asked participants about their perceptions and experiences of harassment over a period of 12 months [12], in the study 27% of women and 16% of men said that they experienced at least one form of harassment during that period of time. The OPN study also identified how women between 16 and 34 years of age felt the most unsafe of any age and sex group when using public transport alone in the dark, while disabled people felt less safe in all settings than non-disabled people [12].

The OPN asked participants about their public experiences, including questions relating to busy public spaces, a quiet street near their homes, and in a park or other open spaces [12]. The OPN study also identified how 15% men and 22% women had experiences of being insulted or yelled at by a stranger in public [12], highlighting the need for more work to enable safer spaces in public. However, domestic violence will often occur in domestic and private settings, making the identification, evidence collection, and reporting processes difficult for the individuals involved due to social isolation [9], or family relations [5]. Private settings such as an individual's home or nursing home have been considered in relation to violent locations, domestic violence cases occurring at home [13] and caregivers in nursing homes consider violence a workplace and safety problem [14]. This requires an approach to data collection that can consider victim privacy, as existing concerns regarding privacy and surveillance [15] could be highlighted.

The private nature of domestic violence has been reported in other studies, even when victims spoke with healthcare professionals [16], indicating the need for user-centred privacy to be at the forefront of any approach to helping vulnerable individuals.

Despite concerns about privacy, victims are willing to come forward for support, within England and Wales, The National Domestic Abuse Helpline was reported to have delivered 50,791 support sessions to individuals for the year ending March 2022 [11], the report also indicates that the number was similar to that of those reported the year previous, not matching the increase in Police reported incidents. Despite the support sessions conducted and the number of police reports [11], some research has suggested that victims of domestic violence or abuse may be underreported. For example, a study that analyses cases of domestic violence against older adults with disabilities reports that the number of victims observed in all cases may be underestimated [17]. A review of the literature on physical elder abuse reports that despite the expectation that medical experts are ideally positioned to detect elder abuse, physicians represent only 2% of the reported cases, among other reasons, due to the uncertainty of diagnostic validity [18, 19]. The Covid-19 pandemic has also been considered to have prevented the reporting of domestic violence cases even as violence continued or increased [20]. Evidence collection from a policing point of view is also difficult, with existing recommendations suggesting regular training programmes and arresting officers collecting defendant and victim statements along with photos of injuries and property damage [21].

Despite the common use of technology in crime prevention in a range of offences, technology is rarely used as a method of evidence collection in domestic abuse or violence cases [21]. Some examples are available, a study in New Zealand and Australia reported that a strategy of using video recordings of the victims initial account and using that recording as courtroom testimony could be an effective use of video recordings in evidence collection [22]. An individual's personal data could also be used, such as: call history log, GPS tracking data, abusive text messages, and metadata on photos [23]. Technology has been used to provide support to young people in the UK, and a report highlighting the risks related to digital-enabled support, such as a perpetrator 'lurking' in the

room, recommends 13 corrective actions in this regard [24]. The use of technology by victims has been considered a risk because perpetrators often use technology against victims [25], however, some articles suggest how technology can instead be used by victims through innovative approaches that support online safety [26]. Some proposals consider how technology can be used to address intimate partner violence by recording and collecting evidence of abuse, reducing feelings of isolation, and giving victims access to essential resources/services [26]. Similarly, designers should consider the broader ethical and sustainability challenges raised in this regard (e.g., [25]). Although a technological solution alone is unlikely to stop domestic violence, the collection of evidence and the improvement in perception of safety is the focus of the work conducted in this thesis, through a focus on privacy preservation and user control of an edge computational device.

1.1.1 Motivation

The motivation to complete the research presented in this thesis is to improve the accessibility of devices that can be used to identify, report, and alert individuals to potential domestic violence. The focus of the work presented in this thesis will investigate the identification of violent language from a computational perspective. Using modern technologies, it might be possible to design and develop systems and algorithms that can support at-risk or concerned individuals. The computational development is motivated by the increasingly pervasive and everyday use of edge devices, such as smart home systems and mobile phones. Furthermore, the effectiveness of computational algorithms and edge device processing power suggests that a privacy-preserving but safe violent language detection system may be possible.

1.1.2 Sustainable Development Goals

The importance of the work is highlighted by the relevance that the work has to multiple UN Sustainable Development Goals (SDGs), through which the UN encourages a shared blueprint for peace and prosperity for people and the planet [27]. The research carried out as part of this project identifies two SDGs as the most likely areas for the work to be impactful and contribute to the wider society.

SDG 5 aims to achieve gender equality and empower all women and girls [28], with targets relating to ending all forms of violence against women and girls in both the public and private spheres. The research carried out in this thesis considers SDG 5 in its design to help prevent and report violence in domestic spaces. The SDG also has a target of 5.b, which is to enhance the use of enabling technology to promote the empowerment of women, which is also one of the goals of the mobile edge application.

SDG 16 aims to promote peace, justice and strong institutions [29], which is related to the recording and reporting aspect of the research project. Should this research be effective, progress can be made on how user-centred privacy can be sustainably included in trustworthy crime prevention technologies. This could help improve the surveillance and security of victims of domestic violence.

1.2 Research Gap

Despite the widespread use of technology in crime prevention [30], limited technological methods have been investigated in the context of preventing domestic violence in spaces that are not online. There are some applications in this regard; for example, Circle of 6 [31] is a mobile application that allows its users to quickly contact a user-defined list of six people and share their current location along with an emergency message. Similar applications, including BrightSky [32] and Silent Beacon [33] are aimed towards victims of domestic abuse or aim to increase safety in the public through a virtual panic button. These applications are user-centred in nature, enabling users to identify their own network of trusted contacts and devices who to send potential panic button messages to. There is a limited crossover between mobile applications and academic literature, with most work focussing on public spaces where individuals feel less safe [12]. There are not many solutions that focus on domestic spaces, which may be due to the global positioning system (GPS) being a key feature of most panic button applications. An identified gap in the literature in this regard is the lack of work that investigates personal safety technologies in a domestic environment.

The prevalence of edge devices such as mobile phones and smart home technologies presents a novel opportunity to support the identification of violent lan-

guage or conversations in domestic environments. Edge computing is a method of collecting, storing, and processing data as close to the network edge as possible, presenting opportunities related to data privacy, bandwidth, and latency [34], which could occur within a domestic setting. Edge computing contrasts to cloud computing by locating data processing without the need for central servers or processing facilities, with most research investigating edge computing usage in the smart home, energy and transportation sectors [35]. Potential edge devices such as mobile phones are already widespread across most age groups (88%) of all UK adults own a mobile phone [36]), presenting a research gap for using a mobile phone as an edge computing domestic violence prevention device. Smart home technologies are also becoming more ubiquitous, with audio-equipped technologies such as smart home speakers (11% of the UK own at least one) increasing in popularity in recent years [37]. Therefore, opportunities to infer signs of domestic on the edge to support user privacy could be potential future innovations [26].

Machine learning models are often used to detect violent language [38], hate speech [39], and harmful content [40] on social networks. However, these examples are often limited to text-based posts on social media sites such as Twitter or pre-recorded video content on sites such as YouTube, with language models that are used on mobile devices often limited to virtual assistants such as Alexa or Siri [41]. Therefore, an identified gap in the literature on machine learning, audio processing, and natural language processing is the combination of text and audio content to detect violent conversations in real time. The existing literature has considered the fusion of multiple text models [42], however, a thorough literature review was unable to identify works that combined audio and text modalities for the detection of violent conversations. Although complex environments such as public spaces often contain large amounts of noise, the private setting of domestic violence suggests a potential gap related to the capture of conversations in relatively quiet locations with a small number of individuals.

Therefore, the research conducted in this thesis will therefore investigate the aforementioned research gap presented in Figure 1.1, notably the lack of research on domestic violence in private spaces, the ubiquitous use of potential edge devices such as mobile phones and smart home technology, and the detection of violent language in conversations. The research gap is a computational solution to detect

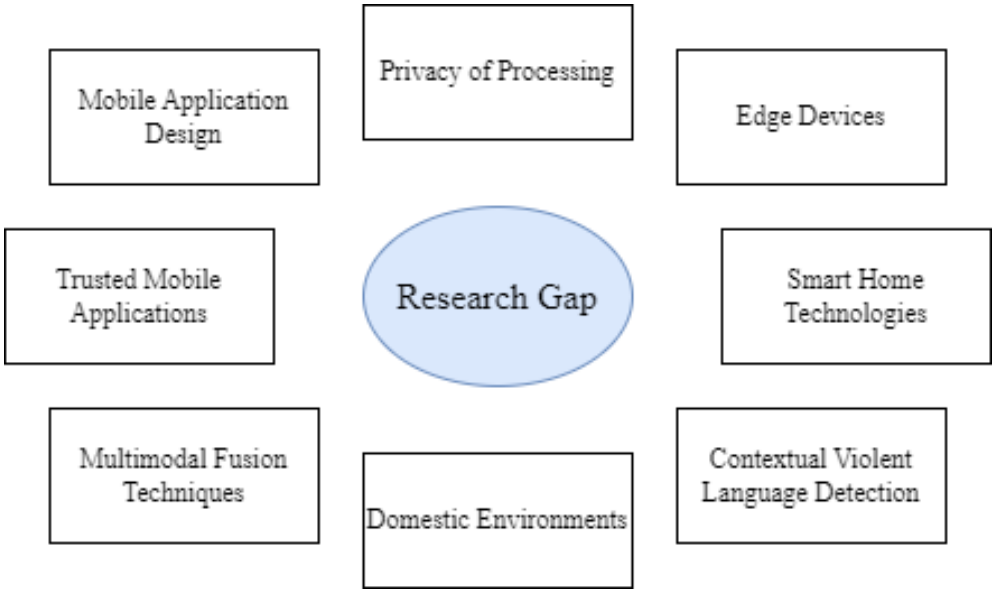

Figure 1.1: Overview of the research gap presented in the thesis, with previous work and literature being used to identify the key research areas of the project. The research gap themes are based on thematic elements such as the domestic environments and technical challenges such as multimodal modal fusion techniques.

violent language in conversations on local edge devices that can preserve privacy while accurately inferring violence.

1.3 Research Question

The identification of the research gap enabled the formulation of a research question, and the research question highlights the technological approach based on grounded theory used, integrating existing literature and knowledge to produce a novel technical solution. The research question is defined as follows:

> Is it possible to infer violence or aggression based on audio and natural language processing on the edge in real-time?

The question was formatted in a way that enabled a three-stage approach to answer the question; this would initially start with the audio and natural language processing models before the data fusion model is developed. Finally, the fusion data model was applied to the edge for the use of mobile phones and

smart home devices. The research question guided the study throughout and enabled constant thought as to the research gap and the potential impact of the completed research.

The purpose of the proposed research is to develop a framework investigating if the real-time detection of violent language is possible in domestic environments for the purpose of crime prevention. This solution could potentially be further developed to send out notifications to the authorities or someone close to the victim. It can also be used to securely collect evidence for court proceedings, however future developments and research is needed alongside stakeholders for this to be further investigated.

1.4 Aim and Objectives

The research question provides an impetus for the study to be carried out and motivates the study choices in relation to data collection, analysis and the method of presentation of the results. Relating to the research question, is the project aims and objectives which define the specific work packages as part of the study. An overall breakdown of the approach used is presented in Figure 1.2.

1.4.1 Aim

The aim of the research is to investigate the fusion and combination of audio and textual content to determine whether violent or aggressive conversations are occurring around an edge computing device. The aim describes the overall research project, without asking a specific question, enabling a contribution to be achieved from the project independently of the outcome. The aim of the research is as follows:

> To analyse the data fusion of multimodal sources to determine if a conversation contains violent or aggressive content on the edge.

In the aim, the multimodal sources refer to audio and text content from recordings and linked transcriptions. The determination of violent or aggressive content refers to the use of binary classification to achieve a result between 0

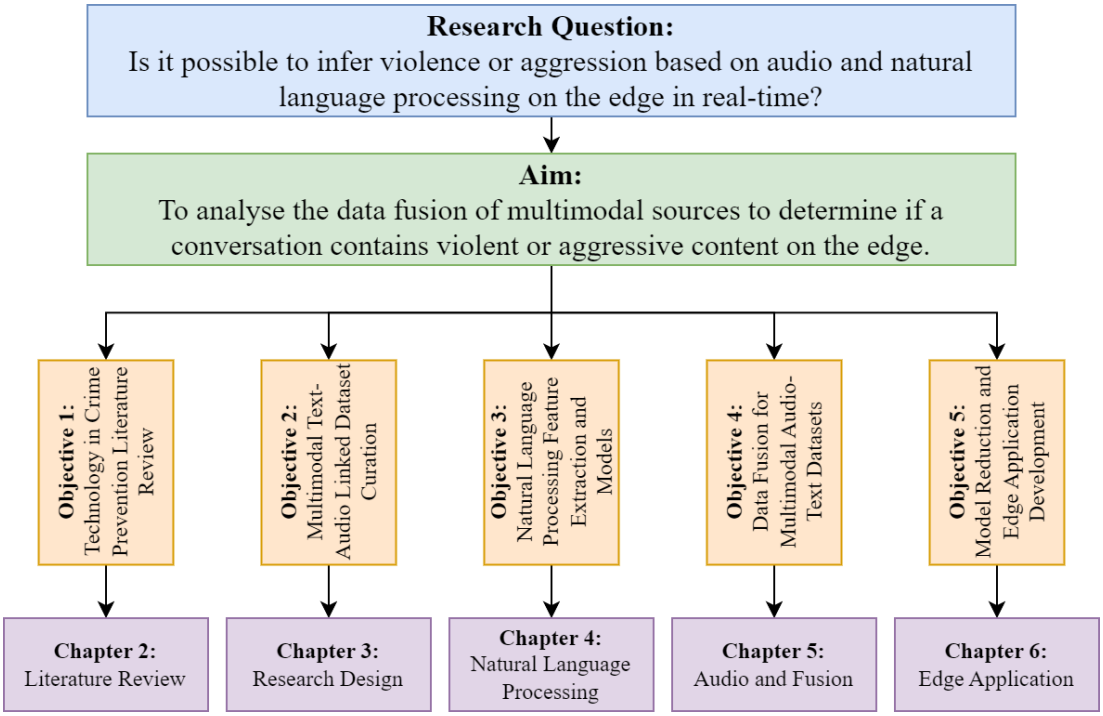

Figure 1.2: The breakdown of the research project, initially presenting how the research question links to the aim of the project. The aim is then linked to each of the objectives, which are linked to the relevant chapter.

and 1. Finally, edge processing refers to the use of mobile phones or smart home technologies, which are the likely devices on which data will be collected and processed.

1.4.2 Objectives

Each objective relates to both the aim and the research question of the project but determines a smaller subset of work that is to be completed as part of the research project. The objectives developed as the study progressed and each result influenced future work and decisions. The objectives of the research project are described as follows:

1. To perform a review and evaluation of the existing literature, sources, and applications relating to the use of technology and Ai in crime prevention.
2. To curate a multimodal dataset of linked audio and text features that can be employed to test and train data fusion models.
3. To investigate the results of natural language processing (NLP) feature extraction, and determine the most effective combination of text-processing models.
4. To develop a data fusion model that combines audio and text modalities to accurately identify violent language from conversations.
5. To perform model reduction on the data fusion model to embed the model on edge devices in real world scenarios.

1.5 Research Challenges

The research project identified several offset challenges that provided motivation for the study, as presented in an overview in Figure 1.3. The figure groups these research challenges based on how each theme impacts the potential implementation of a violent language detection system, these considerations include existing datasets, model complexity, and privacy. Each of the research challenges encouraged the technical and research development that occurred throughout the

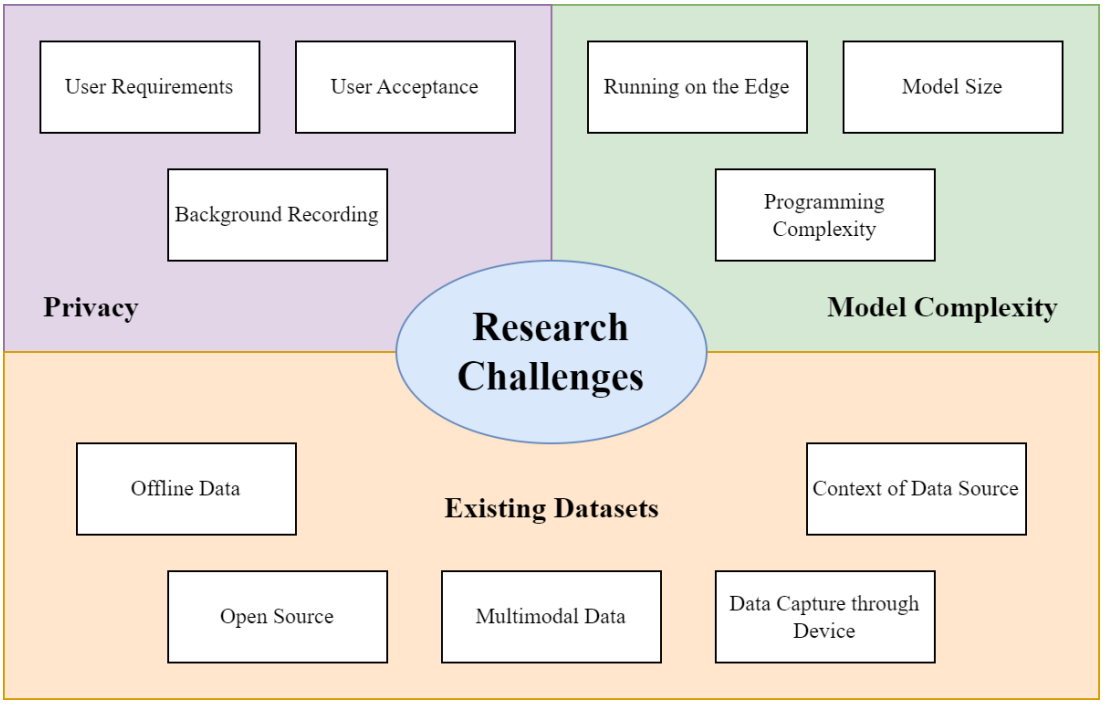

Figure 1.3: An identification of the research challenges that were considered during each stage of the project, split into three main categories. These challenges are identified as themes, with sub-themes also being included in the classification.

study period of the project. The list below highlights some of the main challenges identified during the work and how they were overcome as part of the research.

1. **Existing Datasets**: Despite the interest in online abuse, few studies have considered abuse in a domestic setting. There are several available datasets that focus on abusive or offensive language. For example, The National University of Singapore SMS Corpus; This database contains more than 10,000 abusive SMS messages. The dataset contains the text of the message, as well as audio recordings of the message, read loudly [43]. Another widely used dataset is the "The Toxicity Corpus" [44], which contains a collection of more than 100,000 Wikipedia comments and has been named toxic or abusive language. The dataset includes comments text and audio recordings of comments read loudly. The challenge here was finding a dataset that entails of audio recordings as well as their transcription labelled whether the conversation is violent or normal.

2. **Complex Multimodal Fusion of Audio and Text**: Multimodal data fusion refers to the process of combining data from multiple sources and methods to improve the accuracy and efficiency of the system or process. In multimodal data fusion, different means refer to different types of data, such as text, audio, video, or sensor data. However, when applying multimodal fusion, a large number of computational resources are used, requiring careful considerations with respect to dataset size, data resolutions, missing / incompatible values, data quality, and input and output shapes for each model [45]. Although previous work has combined several text models, such as the fusion of Bidirectional Encoder Representations from Transformers (BERT) and Linguistic Inquiry and Word Count (LIWC) [42], which has not been linked using a multimodal text-audio dataset, which presents a novel challenge for research.

3. **Privacy Concerns**: Victim privacy concerns about the use of technology should be considered throughout the project. Previous research has identified the private nature of domestic violence [16], and should be considered in the design of the study. This is also important due to the potential risks that technology can have on victims [25], however, carefully considered innovation could still be effective if designed with careful consideration [26]. A notable challenge in this regard is the capture of data that could be used as evidence [23] while respecting the privacy of the individual. Consideration should also be given to the comfort levels of the user, as the user should also feel in control of the application, through applying methods that match the user's needs and requirements. The challenge in this regard is to work with offline data using computationally complex models, which will require an effective approach to running models on the edge.

1.6 Research Contributions

The contributions of the research project can be split into four main areas, based on the aim and research question of the project. Contributions include multimodal labelled datasets, NLP fusion, audio-text data fusion, and real-time pro-

cessing occurring on edge devices. A breakdown of each contribution is provided:

1.6.1 Contribution 1: Labelled Multimodal Dataset for Linked Audio and Text Violent Language Detection

The initial research contribution is the curation of a dataset that contains both audio recordings and text transcriptions, linked to unique ID fields. The dataset comprises 1,295 audio files, segmented into 10-second intervals. Each segment is rated on a scale from 1 to 5, based on the presence and intensity of verbal abuse, including violence. These audio files are sourced from various British television series, offering a diverse range of content. The dataset is a unique contribution to the field due to the lack of existing multimodal datasets that are linked correctly or accurately. The dataset is also curated with researcher-identified labelling for whether violence has occurred in the segment's conversation. The dataset is a useful contribution, as it provides a comprehensive and standardised collection of audio and transcription data that researchers could use to study abusive behaviour and its effects. This data could be used to develop and test algorithms or models that are designed to detect or prevent abuse or to better understand the underlying factors that contribute to abusive behaviour. The dataset's components and their attributes are detailed accordingly in Chapter 3.

1.6.2 Contribution 2: Natural language Processing Fusion Model to Detect Violent Language

Although NLP models are common for detecting violent language from social media text data, these models have yet to be compared or analysed in a way that supports high levels of accuracy for the complex contextual nature of human conversation, especially when presented in text format. There have been a small number of attempts at combining multiple NLP models, but these have not been investigated yet in the context of violent language. This contribution provides a comparison of NLP models for detecting violent language, in addition to comparisons of data fusion on disparate NLP models.

1.6.3 Contribution 3: Novel Audio-Text Fusion Model for Detecting Violent Language

The third contribution of this thesis is the audio-text fusion model of linked audio-transcription data to detect violent language. This contribution presents a novel approach to combining multilevel multimodal datasets using binary classification while achieving improved results compared to individual modalities (e.g., text or audio). This approach utilises time-frequency domain features to uncover crucial insights about the audio signal in both time and frequency dimensions. The model also integrates Mel-Frequency Cepstral Coefficients (MFCCs), offering a comprehensive depiction of the signal's attributes which are provided in Chapter 5. The contribution is based on detecting violent language in the context of conversations, which is difficult due to the personal nature of conversational language. Results related to different combinations of multimodal data are also provided, which encourages future work based on improvements in the design of individual models.

1.6.4 Contribution 4: Real-time Edge Processing for Multimodal Violent Language Detection

The final major contribution of the thesis is the design and implementation of a real-time edge processing application, which the previously identified data fusion model could run on. Existing methods of running 'lite' or 'tiny' machine learning models often focus on the design of individual models or modalities. The contribution presented in this thesis runs a complex multilevel audio-text multimodal algorithm on the edge, which is able to achieve the same results as the computer-ran version. This contribution also highlights potential privacy improvements due to the edge nature of the devices, in addition to potential future contributions in trust and safety. The contribution is also demonstrated across two separate devices, a simulated smart home Raspberry Pi and both IOS/Android mobile devices.

1.7 Publications

During the study period, academic impact was considered through the publication of academic manuscripts. The following manuscripts related to the PhD study were published during the research project:

- Woodward, K., Kanjo, E., Anderez, D.O., **Anwar, A.**, Johnson, T. and Hunt, J., 2020, November. DigitalPPE: low cost wearable that acts as a social distancing reminder and contact tracer. In Proceedings of the 18th Conference on Embedded Networked Sensor Systems (pp. 758-759)
- Kanjo, E., Ortega And erez, D., **Anwar, A.**, Al Shami, A. and Williams, J., 2021. CrowdTracing: Overcrowding Clustering and Detection System for Social Distancing.
- **Anwar, A.** and Kanjo, E., 2021, November. Crime Prevention on the Edge: Designing a Crime-Prevention System by Converging Multimodal Sensing with Location-Based Data. In 16th International Conference on Location Based Services (p. 96-100).
- **Anwar, A.** and Kanjo, E., 2023, July. User-Centred Detection of Violent Conversations on Mobile Edge Devices. In International Conference on Human-Computer Interaction (pp. 335-346). Cham: Springer Nature Switzerland.

1.8 Thesis Outline

The thesis is structured into seven chapters, each related to a specific aspect of the research project:

Chapter 2 is a review of the literature on the state of the art in crime prevention technologies, identifying the existing knowledge in the subject area. The chapter also identifies the challenges and opportunities relating to these technologies, which are then used in the determination of the research project.

Chapter 3 outlines the methodology used throughout the thesis related to the development and analysis of the overall project. The chapter also introduces and

explains the process for the dataset curation and describes the dimensionality of the multimodal linked data used as part of the project.

Chapter 4 relates to the research on NLP and text processing for the transcriptions extracted from the audio data. The chapter explains the experimental setup, pre-processing, models used, and the resulting text fusion attempts to determine which model should be used on edge devices.

Chapter 5 presents the work on Inference from Audio, Pre-processing, and feature extraction. The chapter also explains the data fusion process for combining the datasets and models. The chapter describes the experimental setup, processing, challenges, and results of the data fusion attempt in the context of violent language detection.

Chapter 6 explains the real-time solution for processing the data fusion model on edge devices. The chapter describes the challenges, considerations, implementation, and testing of smart home devices, mobile devices, and the overall system.

Finally, Chapter 7 is the discussion and conclusion, which finalises the results of the work, and reports if the research question, aim, and objectives of the project were met, and what future work could be conducted based on these results.

Chapter 2

A State-of-The-Art Review of Technology in Crime Prevention

2.1 Chapter Overview

Recent advances in technology and machine learning provide researchers with opportunities to use these advances to enhance the safety of individuals. This chapter is a state-of-the-art review of the technologies currently available for crime prevention. The high rates of crime in England and Wales, specifically in cases of domestic abuse [11] were the motivation for the review. The review attempts to identify current academic knowledge related to technologies being used for crime prevention. Although there are several concerns regarding the use of technologies for crime prevention [26], the use of such technologies is becoming more ubiquitous due to cost savings and improved infrastructure. Technologies for crime prevention is a subject area growing in interest, as it can be applied to both individuals and organisational entities. For example, a closed-circuit television (CCTV) network [46] can be used for individual domestic security, public spaces, and corporate environments. Other security solutions are often used in specific scenarios. For example, smart home technologies are often only used in domestic environments [37], while facial recognition is often only used by police or public/private sector security solutions [47]. Similarly, software or algorithm solutions are also identified in the chapter, such as social media crime preven-

tion [48] or the use of machine learning technologies [49]. Each of the potential implementation scenarios, along with known crime prevention technologies, is reviewed as part of this chapter. Additional literature has been attached in Appendix A which reviews audio-visual technologies and ubiquitous sensing. The remainder of this chapter provides a state-of-the-art review of crime prevention technologies with a focus on data science, software, and mobile technologies.

2.2 Background

Criminal activity is a prevalent issue in contemporary culture and society, and most countries face intolerable levels of crime [50]. Technological innovation has been one of the main driving forces leading to the continuous improvement of crime control and crime prevention strategies. For example, GPS tracking has been used to track offenders' movements [51], Bluetooth has been used to monitor and track large crowds at public events [52], and electronic monitoring has been used to track past offenders and prevent recidivism [53]. Therefore, technology has been used by police and national organisations to prevent crime from occurring and to encourage the perception of safety to the general public. Despite this, technologies and solutions are often specific to the individual context, for example, while GPS [51] and electronic monitoring (EM) [53] are used to track offenders, these technologies are rarely used to monitor and protect potential victims, due to concerns such as privacy [54]. The growth and ubiquitous nature of more novel technologies has also been considered in the academic literature, with authors considering the use of machine learning [55] or artificial intelligence (AI) [56] in crime prevention technologies. In addition to the use of technologies for traditional crimes, the use of technology has presented new crimes that governments should consider, such as those that occur on social media, including hate speech [57] and offensive language [58]. Similar considerations should be taken with regard to cyber crimes, including phishing [59], online fraud [60], and digital impersonation [61]. Therefore, it is necessary for researchers to consider the crimes that occur and if technological solutions can be used to support the prevention of such activities.

First, it is necessary to define the crime and the main criminal activities that

occur. Crime is a broad term used to describe a wide range of criminal activities, and the academic literature and official statistics indicate significant variation in both risk and reporting rates between the different types of offence. For example, according to the Office for National Statistics (ONS) [62], the most common type of crime against a person in the UK in 2021 was violence against a person, which represented 23% of all reported crimes. This category includes crimes such as common assault, harassment, and stalking. The second most common type of crime against a person was sexual offences, which represented 11% of all crimes [62]. However, reported crime statistics do not consider an individual's perception regarding their safety, for example, according to the ONS individuals felt less safe walking alone in all settings after dark than during the day [12], which while linked to personal safety and crime prevention, is not often reported as crime statistics. Crime prevention also includes communication with communities, through the provision of expertise and advice. For example, Police UK provides online advice on potential criminal activities such as spiking, drugs, and fraud [63]. In this article are recommendations on technologies for the domestic environment, such as marking a property with ultraviolet light and registering the items in an approved database to improve the chances of recovering stolen goods [64]. The crime statistics and crime prevention advice highlight the importance of crime prevention, but also the necessary improvements that could be made through the widespread use of modern technologies.

Previous literature reviews and surveys related to crime prevention and technology have been conducted, most of which are related to specific contextual implementations (e.g., NLP and crime prevention [39]). However, it is necessary to conduct a review that investigates the use of technology in crime prevention as a whole, to enable the detection of novel methods of preventing crime. Articles have previously identified the opportunity that technology and innovation have in relation to criminal activities [65], however, the article also identified how researchers must consider the weakness that purely technological approaches can have in crime prevention. A systematic review on technology and crime prevention has also been conducted, focused on specific crime prevention journals [66], this review highlights common themes in the literature, including artificial intelligence and information technologies. The review [66] uses a systematic approach

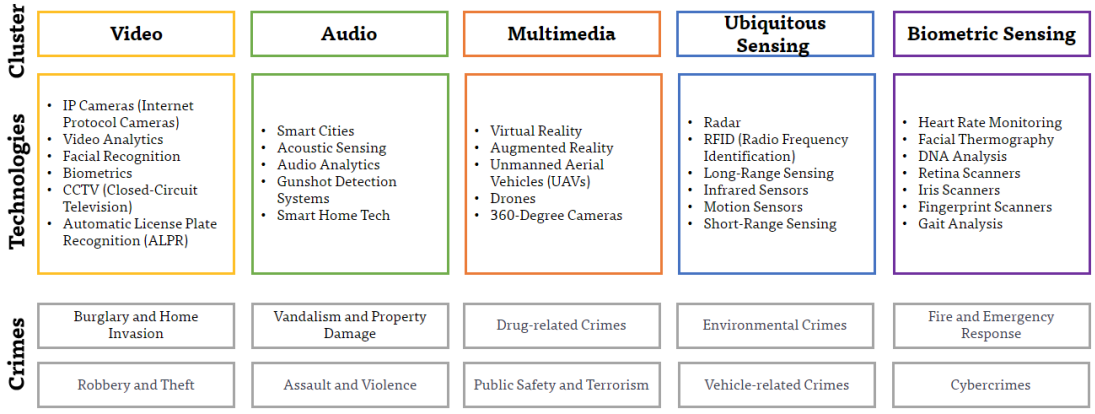

Figure 2.1: A classification of the literature that was reviewed during the research project, classified into main themes, sub-themes, and examples crimes where the technology could be used. The classification is based on a narrative review, with the themes being identified during this process.

that means that novel methods with limited research interest could be ignored or not included in the systematic process. Reviews have also been conducted related to specific contexts of technology in crime prevention, such as rural situations, where a review of the literature [67] identified how despite an increase in use in the last decade, technology is not common to prevent rural crimes. This indicates the need for technologies to be more accessible and shared more frequently to be applicable to the wider contexts of crime. Similar work has also been carried out related to specific innovations, for example artificial intelligence, which has been considered in the context of cyber crimes [68] in the previous literature. The nature of the identified existing work highlights the need for the review performed in this chapter, which considers the role of technology, in non-specific criminal contexts, providing motivation for future work.

2.3 Literature Search Method

The literature reviewed within this section was curated using a manual search due to the disparate nature of the research area. For this purpose, a state-of-the-art non-systematic narrative review [69] was performed; this was used as a review method to identify current knowledge within technology in crime prevention and

to provide an impetus for research carried out within the remainder of this thesis. The initial plan was to conduct a systematic survey of articles or an integrative review [70], however, limitations in the terminology of naming technologies and the large number of collected irrelevant articles meant that a narrative approach was taken instead.

The potential limitations of this literature search methodology are the lack of reproducibility and transparency as part of the review; for example, a scoping or integrative review [69] will include inclusion criteria as part of the work and structured search queries to support the discovery of articles for other authors. The lack of an existing structure or literature review in this regard restricted the review to being narrative and state-of-the-art in nature. Identified as part of this review is the potential for a systematic review to be conducted in the future, based on the knowledge formed from this state-of-the-art review, where topics and search methodologies can be based upon those identified as part of this work.

Figure 2.1 presents the themes identified during the literature review process. The themes include main themes such as audio-visual technologies, subthemes such as smart home technologies, or acoustic sensing. The figure also provides references to examples of crimes considered within the research papers, and these provide context as to how and where the prevention technologies are used.

2.4 Data Science

Data science methods are some of the most common techniques to be linked with crime prevention applications, due to well-known reports of predictive policing being implemented into police tool kits [71]. Data science methods refer to the use of techniques such as big data [72], or machine learning [73] to effectively prevent crime using novel technologies and applications. Data science methods are often based on algorithms and mathematical concepts. Despite this, concerns have been raised about the potential bias in such systems [74]. Limitations can present potential ethical implications [75] and racial bias [76] in data science methods, and therefore researchers must consider this in their work.

2.4.1 Machine Learning and Deep Learning

Machine learning is a data analysis method that automates the development of analytical models. It is a field of AI based on the idea that a system can learn from data, identify patterns, and make decisions without human intervention [77]. There are different types of machine learning; such as Artificial Neural Networks (ANNs), CNNs, Deep Learning, Spatial-Temporal Models, and Empirical Models. While CNNs are a particular kind of ANN that is tailored for processing grid-like data, like images, ANNs are a more general term that refers to any kind of neural network. Deep learning makes use of both ANNs and CNNs, which are neural networks (NNs) with many layers [78]. These techniques can be useful and has been used for crime prevention tools, as they use historical crime data to predict where and when crimes are likely to occur, identify new trends in crime, improve efficiency, use facial recognition to identify suspects [79], analyse the feelings of social media [80], and provide real-time crime warnings [81]. While the majority of these techniques are already identified within this review, this section and the included sub-sections will focus on the use of the algorithms as opposed to the contextual use-cases.

Deep learning is a method of machine learning that excels at managing unstructured data and extracts features from it using layered models. It has gained importance as data has increased and computer hardware has advanced [82]. One of the key technologies used in driverless cars [83] allows them to recognise a stop sign or tell a pedestrian from a lamppost. Large amounts of data are used to train deep learning models, which can be applied to a variety of tasks, including decision making [84], speech and image recognition [85], and NLP [55]. These techniques are particularly effective for tasks that involve the analysis of complex and unstructured data, such as images, videos, and audio [86].

This study [74] presents a data-driven strategy for analysing crime data and predicting the development of crime hotspots in Taiwan using machine mining algorithms based on the “broken windows” theory and geographic analysis. The Deep Learning algorithm is employed and has been shown to outperform other approaches such as Random Forest and Naive Bayes. Using data with multiple time scales improves model performance. The results are confirmed by mapping

possible locations of the crimes. Another study [87] found that deep learning-based models can be used to detect cyberbullying on numerous social media sites. Similarly, research [88] has explored how an intelligent video surveillance system that actively monitors in real time without human intervention using deep learning technology can be used.

A type of deep learning algorithm is a recurrent neural network (RNN). RNNs are intended to handle sequential data, such as time series data or text, by retaining an internal memory or state that allows the network to learn and anticipate based on past input. RNNs are very good for jobs that involve sequential information and are utilised in a wide range of applications, including NLP, audio recognition, and time series prediction.

2.4.1.1 Artificial Neural Networks

ANNs are modelled after the structure and operation of the human brain and are a fundamental building block of deep learning. It is made up of a network of interconnected "neurons" that transmit and process data between layers. Each neuron receives information from another neuron, processes the information mathematically, and then sends the results to another neuron or output layer. The "weight" set of neurons, which is adjusted throughout the learning process to increase network accuracy, represents the strength of the connections between neurons.

An ANN (and often, more conventional NNs) consist of an input layer made up of neurons (or nodes, units), one or more hidden layers, or even three hidden layers with output neurons as the final layer [89]. The lines connecting neurons are also shown in Figure 2.2, which depicts a typical architecture. A weight, which is a numerical value, is assigned to each connection. An example of this in mathematical form is presented below:

$$h_i = \sigma\left(\sum_{j=1}^{N} V_{ij} x_j + T_i^{hid}\right). \tag{2.1}$$

Where the output of the hidden layer neuron is h_i, N is the number of input neurons, V_{ij} are the weights, x_j are the inputs to the input neurons, and T_i^{hid} is the threshold. the hidden neurons in terms.

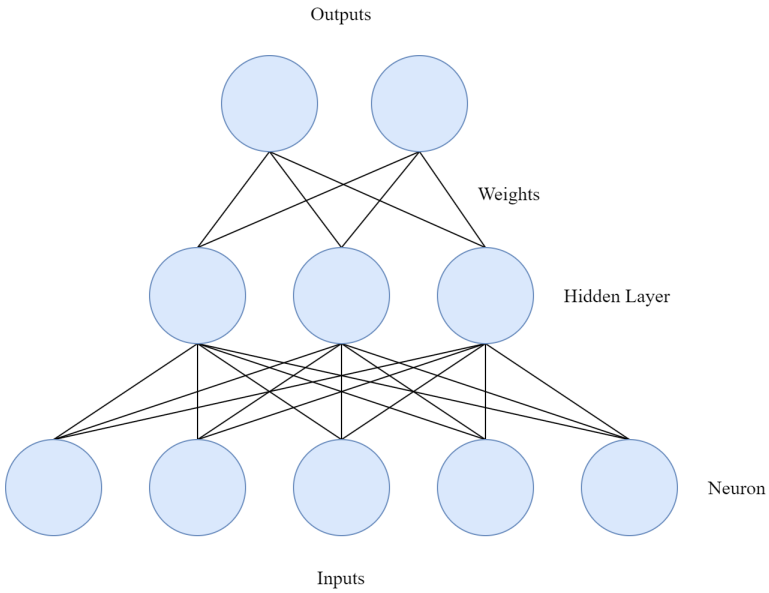

Figure 2.2: A demonstration figure representing a typical neural network architecture. The neuron represents the individual nodes of the network, the weight is the weightings of each node, the hidden layer represents the layer between input and output. The input and output values are also presented, being the start and end of the architecture.

In addition to adding non-linearity to the NN, the activation function aims to constrain the neuron's value to prevent the NN from being paralysed by divergent neurons. The equation below represents the sigmoid function denoted by $\sigma(u)$. The function takes an input u and maps it to an output value between 0 and 1. It is defined as:

$$\sigma(u) = \frac{1}{1 + \exp(-u)}. \tag{2.2}$$

The sigmoid function is commonly used as an activation function in NNs. It introduces non-linearity to the network, allowing it to model complex relationships between inputs and outputs. The output of the sigmoid function tends to 1 for large positive values of u, tends to 0 for large negative values of u, and approaches 0.5 as u approaches 0. This property makes it useful for tasks like binary classification, where the output needs to be in the range of 0 to 1, representing the probability of a particular class.

An overview of NN applications has been conducted in real-world settings [90], and current trends in ANN research are presented along with a taxonomy of ANN. The study covers a wide range of ANN applications, including computer science, engineering, medicine, environmental protection, agriculture, mining, technology, climate, business, and the arts. It shows that ANNs with feedforward and feedback propagation work well in their intended applications. To achieve better performance, the study recommends that future research concentrates on combining feedforward and feedback propagation ANN models into a network-wide application.

To anticipate and forecast the emergence of drug hotspot areas, this article [91] discusses the development of an early warning system that makes use of GIS and artificial NNs. The system uses a pre-trained NN to process geographic information system (GIS)-based data, and it then shows the results on a map by emphasising regions where there are likely to be a lot of drug-related 911 calls. By anticipating criminal activity, this system seeks to assist proactive law enforcement efforts.

The Artificial Neural Network-Artificial Bee Colony (ANN-ABC) algorithm, which is the hybrid crime classification model proposed in the literature [92], combines the ANN and Artificial Bee Colony (ABC) algorithms. The objective is to use ABC as a learning tool for ANN in order to avoid the local optima problem and generate more meaningful results. The Communities and Crime dataset is used to apply ANN-ABC to predict "Crime Categories" from the UCI machine learning repository. When the results were contrasted with those of other classification algorithms, it was discovered that ANN-ABC outperformed them with an accuracy of 86.68% and an average improvement of 7.5%. Therefore, ANNs can be applied to a variety of tasks, including voice, image, and NLP. These are especially beneficial when handling complex nonlinear relationships in data.

2.4.1.2 Convolutional Neural Networks

CNNs are a particular kind of ANN architecture, with a precise yet straightforward architecture, CNNs offer a streamlined way to get started with ANNs and

are primarily used to tackle challenging image-driven pattern recognition tasks. CNNs are made to process data with a grid-like topology, like an image, by extracting features and reducing the dimensionality of the input data using a series of convolutional layers and pooling layers. They are commonly employed in image classification, object detection, and computer vision [93]. An example of a basic CNN in mathematical form is presented as follows:

$$\text{Convolution: } \boldsymbol{z} = \boldsymbol{w} * \boldsymbol{x} + b \tag{2.3}$$

In the convolution operation, a filter or kernel $\boldsymbol{w}$ is applied to the input $\boldsymbol{x}$ to obtain feature maps $\boldsymbol{z}$. The $+ b$ represents a bias term added to each element of the output feature maps.

$$\text{Activation: } \boldsymbol{a} = \text{ReLU}(\boldsymbol{z}) \tag{2.4}$$

The ReLU (Rectified Linear Unit) activation function is applied element-wise to the feature maps $\boldsymbol{z}$. ReLU sets all negative values to zero and keeps the positive values unchanged, introducing non-linearity to the network.

$$\text{Pooling: } \boldsymbol{s} = \text{maxpool}(\boldsymbol{a}) \tag{2.5}$$

In the max-pooling operation, the maximum value is selected within nonoverlapping regions of the input $\boldsymbol{a}$ to create the pooled output $\boldsymbol{s}$. This reduces the spatial dimensions of the feature maps and retains the most important features.

CNNs have seen significant advances in deep learning in recent years, due to the expanding availability of annotated data and improvements in hardware have contributed to these developments. In a variety of tasks, such as visual recognition [94], speech recognition [95], and NLP [96], CNNs have produced state-of-the-art results after extensive study. In an article on recent developments in CNN research [97], the authors offer a comprehensive overview including improvements to layer design, activation functions, loss functions, regularisation, optimisation,

and quick computation. In addition, the article [97] covers a variety of CNN applications in speech, NLP, and computer vision.

For example, the method described in an article on human activity classification [98] uses motion capture data, time-frequency analysis, and a complex value convolutional neural network (CV-CNN) to classify activities using radar data and deep learning. Through the analysis of motion trajectories and radar echoes obtained from the motion capture data, micro-Doppler characteristics are discovered, and a sample database is created. The data is used to train CV-CNN, and the results show a high classification accuracy of 99.11% [98]. Furthermore, the method [98] exhibits improved performance compared to conventional CNN methods in a variety of training sample proportions and signal-to-noise ratios.

Another study [99] examines how to identify suspicious behaviour that could occur before a crime is committed in order to address the issue of shoplifting crimes in surveillance videos. Instead of identifying the crime itself, the suggested method uses a 3D CNN model as a video feature extractor to find sections of a video that are highly likely to contain a shoplifting incident. This gives surveillance personnel more opportunities to intervene and stop crime. The model has been tested on a dataset consisting of daily action and theft samples, and the results show that it correctly predicts the imminent commission of a crime in 75% cases. This strategy shows promise in addressing the issue of surveillance personnel's inability to process massive amounts of data in real time, thereby lowering the losses brought on by shoplifting crimes. Object detection, face recognition, and image segmentation are just a few of the tasks that CNNs are effective for handling in image and video processing. CNNs have gained popularity for many deep learning applications as a result of their success in image and video processing tasks.

2.4.1.3 Spatial-Temporal Models

To understand the surroundings and the reality of the world, a spatial analysis is required. The challenging issue of location analysis can also be solved by spatial analysis. This enables a process to determine whether patterns persist over time and if they are and to identify any unusual patterns with the aid of a temporal

understanding of the data [100]. A spatial-temporal model arises when data are collected over time and space, with at least one spatial and one temporal property. Events in spatio-temporal datasets describe spatial and temporal phenomena that exist at a certain time t and location x [101].

A systematic review of the literature on the detection and prediction techniques of crime points used in this study [102] reveals a lack of detailed studies in this field. The review consists of 49 studies and discusses the use of deep learning and time series analysis, as well as the incorporation of spatial and temporal information into crime datasets. The application of deep learning techniques for spatio-temporal data mining (STDM) has made significant strides recently, and this article [103] provides a detailed review of these developments. Deep learning models such as RNN and CNN are frequently used in STDM tasks as a result of the increasing availability of spatio-temporal data, which renders traditional data mining methods inadequate.

Another paper [103] has an effective real-time crime prediction technique where the authors adapt the spatial temporal residual network and use a proper representation of crime data to forecast how crimes will be distributed over Los Angeles at the scale of hours in parcels the size of neighbourhoods. A ternarisation technique is used to address resource consumption issues for the deployment of the proposed model in the real world. The proposed model is improved in terms of accuracy compared to existing approaches. Similarly, the literature [104] uses machine learning for grid-based crime prediction, embracing the idea of a criminal environment through the application of spatial-temporal attributes based on 84 different sources of geographic data. Deep NNs were discovered to be the best model, outperforming other techniques. According to the findings, the use of geographic characteristics increases the model's performance and capacity to explain crime relocation.

2.4.1.4 Empirical Models

Empirical models, which are used to describe real-world phenomena and are developed from data and observations, are created by analysing the data and looking for patterns and relationships. They come in a variety of shapes and sizes, includ-

ing mathematical, statistical, and machine learning models, and can be applied to make predictions, run simulations, and gain understanding of intricate processes. They are improved upon and tested using additional data rather than presumptions [105].

Machine learning uses a variety of techniques, including empirical methods and ensemble methods. Ensemble methods aim to improve performance over empirical methods by training multiple models and combining their predictions into a final decision. Empirical methods involve training a single model on a dataset to make predictions. An empirical approach to crime prevention that makes use of machine learning is predictive policing. To effectively allocate police resources for crime prevention, it uses historical crime data and machine learning algorithms to predict where and when crimes are likely to occur in the future. "Hot spot policing," [106], which focusses on localities with high crime rates, is an example. To increase the precision of their predictions, predictive policing systems can also use additional data, such as weather, social media, and demographic data. Comparison of the crime rate in areas where predictions were made and the crime rate in areas where other methods were used to allocate resources can be used to evaluate.

Existing literature [107] has explored three ensemble machine learning algorithms and tested the approaches for their viability in predicting air overpressure caused by explosions in open-pit mines. The study used data from 146 blast events, with 20% of the data used to verify the accuracy of the models and 80% of the data used for the development of the models. The results demonstrated that the ensemble models outperformed the empirical approach of a single model, and the Cubist algorithm showed the best performance. The study also identified crucial elements for predictive models. Another example proposed is the the Smart MCBT tool [108] that makes use of cutting-edge data analytics to support policy formulation, particularly in the area of crime prevention. In addition to pointing out that this tool is still being developed, the article also discusses some of its potential advantages and difficulties. The concepts and methodology presented in the article will be tested and evaluated in subsequent work. Not just for crime prevention, but also in a variety of policy contexts.

2.4.2 Natural Language Processing

NLP is a subfield of AI that seeks to enable computers to understand, analyse, and generate human language. These technologies can play an essential role in helping capture important evidence in extreme cases of domestic incidents to protect victims. There has been an increase in interest in detecting violent language with the aim of making the physical and digital worlds safer through literature investigating social media sentiment analysis and cyberbullying [109].

Text analysis has the potential to play a role in the detection of violence by identifying patterns and indicators of violent behaviour in written and spoken language. For example, certain words or phrases commonly associated with violence, such as threats or aggression, could be flagged by a text or audio to text analysis system. However, language interpretation is complex and can be influenced by many factors, such as context [110]. NLP technology has recently emerged in the world of research in different scenarios; although it is not directly related to crime prevention, it greatly assists white collar crimes. However, a common use of NLP in has been to detect increasing cybercrime on social networks [111].

BERT [112], is a state-of-the-art language processing model developed by Google. It is capable of learning contextual relations between words in a text, making it particularly effective in natural language tasks such as sentiment analysis and text classification. In the context of crime prevention, BERT has been used to analyse large volumes of text data, such as social media posts [113] or online news articles [114], to identify sentiment or fake news. Similar studies have explored the use of NLP to identify and analyse patterns in large volumes of text data, such as social media posts, news articles, and police reports. For example, a study applied topic modelling and sentiment analysis to Twitter data [115] to identify communities and individuals at risk of violence. Another study used named entity recognition and word embedding to analyse news articles and identify factors associated with gun violence [116].

To extract knowledge to support criminal investigations, previous work [117] describes a modular approach to processing data gathered from various sources. The suggested platform offers cutting-edge methods and effective parts to handle multimedia data in a scalable and distributed manner, enabling law enforcement

organisations to analyse and multidimensionally visualise criminal data in a single, secure location.

Another study [73] describes how NLP is used to analyse interview content to demonstrate the effectiveness of linguistic patterns in identifying signals of school violence. The study found that the linguistic characteristics outperformed the household data of the subject to predict the risk of violence. The selection of features revealed numerous warning markers that could be useful for individualised interventions, and the best-performing classifier was able to accurately assess the risk levels of the subjects.

Overall, these studies suggest that NLP can be a useful tool for identifying and analysing patterns and trends in text data that may be relevant for crime prevention. However, more research is needed to assess the effectiveness and limitations of NLP in this context and explore potential ethical and privacy concerns.

2.4.2.1 Sentiment Analysis

Most research surrounding sentiment analysis is focused on detecting negative and positive sentiments in data collected from social media platforms such as Facebook, Instagram, and Twitter [118, 119]. In recent years, sentiment analysis has gained immense recognition in academic literature, as the analysis of an individual's emotional state and its dynamics can provide research with cues that could be used for predicting personality and speech patterns. These predictions can be used as a form of violence detection [116] across a range of scenarios.

Sentiment analysis has become more mature in the recent decade [120] and the most widely deployed classification techniques were support vector machines (SVM), Naive Bayes, and Maximum Entropy, which are based on the word bag model. However, the word bag model disregards the sequence of words in a sentence, which can have a significant effect on the meaning of the sentence and change the sentiment, as discussed in a survey conducted on distinguishing between facts and opinions using NLP [121].

The detection of violent and offensive language is conventionally classified into specific types; such as the detection of bullying as seen in [122], identification of aggression [58], and the identification of hate speech [123]. NLP has been

applied to a great extent in studies surrounded by sentiment analysis to take advantage of the syntactic lexical features of phrases and sentences to detect offensive language [124]. As the NLP literature grows and becomes increasingly popular for automatic detection of hate speech and abusive language, common patterns can be seen. Initially, data goes through pre-processing to gain useful insights and to clean up the text from any irrelevant information using methods such as removing Punctuation, Stopwords, Tokenisation, parts of Speech Tagging, and Lemmatisation [125].

2.4.2.2 Social Media

Social media can be a useful tool for preventing crime in several ways. First, law enforcement agencies and community organisations can use social networks to educate the public about crime prevention strategies and provide updates on crime in the area. This can help people be more aware of potential dangers and take steps to protect themselves and their property [126]. Second, social media can be used to encourage community participation in crime prevention efforts. For example, people can use social media to share information about suspicious activities or potential threats, which can help law enforcement agencies identify and prevent crimes before they occur. Outside of NLP, social media can be used to mobilise community members to participate in neighbourhood watch programmes or other initiatives that promote public safety [127].

According to a study [128] on the perception of using social media to prevent crime in communities, members of specific groups, such as African Americans, those living in high-risk areas, and those who have little faith in their local police, are less likely to think that social media is a useful tool. The potential misuse of information, trolling, and being considered a snitch were among the concerns expressed. However, participants also saw social networks as a tool that can support in-person efforts and help build relationships and share information. Taking into account social and historical contexts, the study brought attention to the difficulty of using social networks to prevent crime in the neighbourhood.

This [129] paper proposes ATHENA, an automated system that uses data mining and social media to stop criminal activity before, during and after major

crises. The system coordinates police and law enforcement efforts by gathering and examining social media data about crises. The paper makes the case that these techniques should be used to address problems caused by crises and emphasises the novelness of the system suggested in this regard.

Another [130] study looked into how 122 police officers in Lagos, Nigeria, used social media for police and crime prevention. The study found that most police officers had a favourable attitude toward the use of social networks for police and crime prevention. Most officers (77.2%) did not have training in the use of social networks, despite that collecting information was the main reason why social networks were used.

More discussions have been conducted on the use of social media data and technology for various purposes related to public health and safety, as NLP has grown in the social media space. This article [131] first described a crime investigation tool that uses real-time data from social network services such as Twitter to support crime analysis and visualisation, with a prototype created for the San Francisco area. Another article [132] discussed the lack of research on social media use in rural areas, but made the case that social media use could help reduce crime. Finally, similar literature [133] focusses on the moral and legal implications of using technology-based suicide ideation detection mechanisms on social media platforms as well as the importance of suicide prevention programmes in light of the effects of the coronavirus pandemic on mental health. In general, studies point out the potential advantages of using social media data and technology to improve public health and safety.

2.5 Software and Mobile Applications

Information technology methods are used in everyday police and crime prevention techniques. These include modern technologies and software applications that are currently used by Police forces. Information technology solutions include mobile applications [134], mapping systems [135], user interfaces [136], and risk assessment technologies [73]. These technologies enable methods to engage with data collected from crime or improve potential methods of reporting crimes to the police [134]. Therefore, the content of this section explores software-based

technologies and mobile crime prevention applications, providing a split between user safety and wider scale systems.

2.5.1 Software Applications

Policing services can use software applications to rapidly process, analyse, or present data that officers can use in the field to improve effectiveness. Due to the wide use of Web 2.0 technologies [137], ease of modern processing techniques [74], and cloud-based nature of most systems [138], software applications present novel methods of engaging with and using crime data [139] to assist in crime prevention. As part of this review, two major software applications are considered, as similar aspects are reported throughout this review, including crime mapping and risk assessment.

2.5.1.1 Crime Mapping

Crime mapping, also known as hotspot policing [140], is a method of using crime data to produce maps. Crime mapping is the process of performing spatial analysis to map, visualise, and analyse crime patterns [141]. Crime mapping can be performed by collecting existing police response and resource data, and applying these to information systems or machine learning algorithms. Using these technologies allows the identification of crime hotspots in conjunction with other trends and patterns of crime to be identified [142]. Such technologies and techniques have also been referred to as location technologies [143], because location-based data is a key element in the development of such systems. This information can then be used to optimise the location of human or/and technological resources. As various works suggest, the identification of hot spots is an effective software-based technology to optimise the use of resources and ultimately prevent crime [144]. Existing work has, for example, compared the effectiveness of frequency and length of hotspot patrols in a randomised trial [145]. Other developments have attempted to map the dynamics of crime through interactive visualisations that consider seasonality, crime types, regions of interest, and crime patterns [146]. The authors of this work provide two case studies, including a vehicle robbery in São Paulo City to evaluate a crime pattern theory of road in-

frastructure linked to vehicle crime and passerby robbery in São Carlos related to the effects of urban infrastructure on crime predictions, before receiving feedback on the proposed solution from experts, which is a common theme in the crime mapping literature.

A common method of presenting such data is through GIS, which presents spatial information and allows analysis to be performed. The role of GIS in crime mapping is increasing [135], due to the effectiveness of computationally supported decision making and the presentation of the analysis performed in GIS. Most work using GIS enables multiple mapping methods to be performed, such as an article investigating a web-based GIS for crime mapping and decision support [147], which includes: Choropleth mapping, grid mapping, spatial ellipse mapping, and kernel density mapping. crime mapping software can also allow members of the public to better report crime in their local area; for example, a crime mapping report mobile application that was developed allows residents or visitors to an area to report crimes, criminal behaviour, or potential problems using location-based services [134], which can further improve the effectiveness of collected data and future mapping options. This enables the concept of citizen science to be implemented into crime mapping applications, which can supplement authoritive data [148], and where citizens are used as a sensor to collect data on their locality.

2.5.1.2 Risk Assessment

Risk assessment is another key information-based technology for crime prevention. Risk assessment is used to assess the risk of recommitting a crime by offenders under correctional control. According to the survey conducted in [149], most serious crimes are committed by a small fraction of people during the first months of probation parole. Risk assessment tools make use of predictive models to identify such a subgroup of people so that appropriate surveillance / supervision is given to those cases. Similarly, information technology is used to identify the probability that a terrorist attack or serious violent event will occur in certain places, including schools, airports, or train stations, among others [150]. Another application in which information technology has been adopted to prevent crime is in the development of computer software to track individual interactions on

various social networks [151]. The monitoring of such suspect's interactions is then used to identify abnormal behaviours which can potentially be related to crime intentions.

2.5.2 Mobile Crime Prevention

Mobile crime prevention relates to the use of mobile technologies to prevent crime, mobile technologies are ubiquitous in the modern world [36]. Mobile devices enable safety and security technologies to be portable with an individual while also recommending potential safety improvements in a day-to-day routine. Mobile applications often focus on reporting technologies, such as panic buttons or GPS trackers [152], however, more modern devices such as the iPhone include extended emergency features such as crash detection and satellite calling [1].

2.5.2.1 Mobile Features for Emergencies

There are several emergency mobile features available on the iPhone. These features are designed to provide quick and easy access to emergency services and other important information in the event of an emergency situation, as presented inf Figure 2.3 [1]. One of the main emergency features of the iPhone is the ability to call emergency services by pressing and holding the power button. This feature is available on all iPhone models and allows users to call emergency services, such as police or ambulance, without unlocking the phone or accessing other applications as presented by options in Figure 2.5 and the call in Figure 2.6 [1].

Another key emergency feature of the iPhone is the Medical ID. This feature allows users to create a digital medical profile that includes important information, such as allergies, medications, and emergency contacts as presented in Figure 2.4. This information can be accessed by emergency services, even if the phone is locked, by pressing the power button and then selecting "Medical ID" from the Emergency SOS menu presented in Figure 2.3 [1].

In addition to these core features, the iPhone also includes several other features and apps that can be useful in emergency situations. For example, the Find My app allows users to locate their lost or stolen iPhone [153], and the Weather

Figure 2.3: Screenshot of the Apple iPhone emergency call and medical ID options when the close button is selected. The options enable emergency calls or medical information to be accessed without knowing the password for the device. (Image credits: Apple [1]).

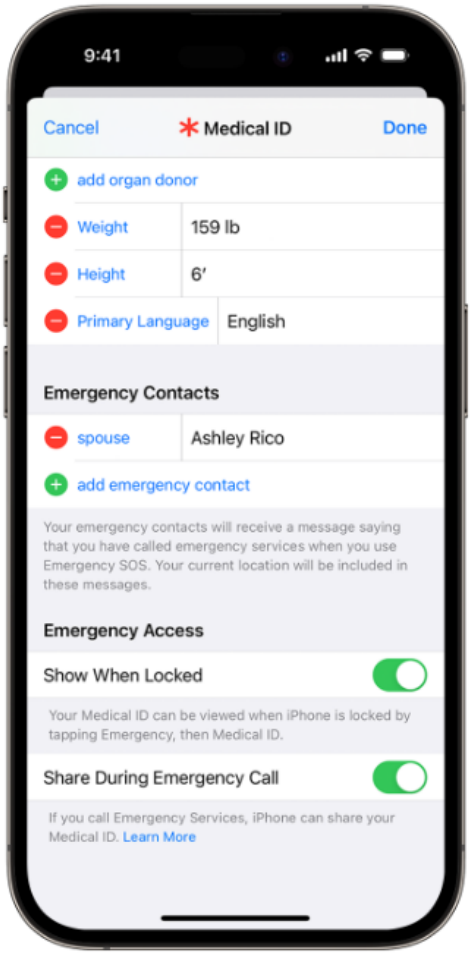

Figure 2.4: Screenshot of the Apple iPhone Medical ID screen, which presents medical information about the device user. Details include: languages, organ donation, height, weight, and custom emergency contacts. (Image credits: Apple [1]).

app provides real-time information on severe weather conditions. Overall, the iPhone provides a variety of mobile features for emergencies that can help users access emergency services quickly and easily and other important information in the event of an emergency.

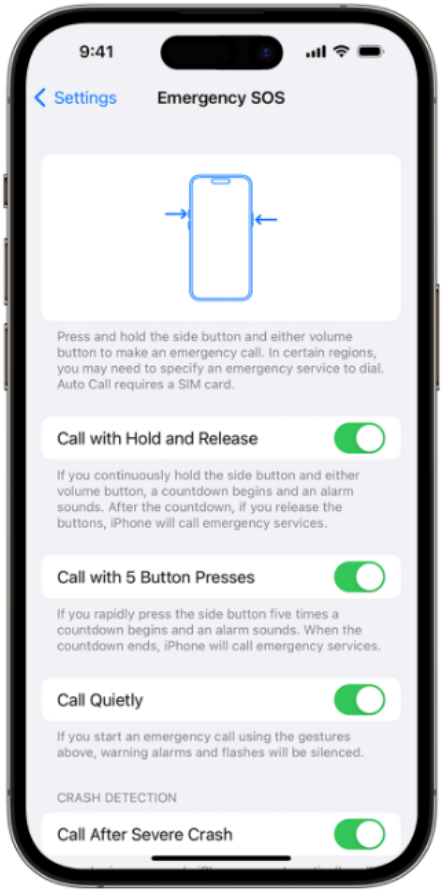

Figure 2.5: Screenshot presenting the user options for the iPhone emergency settings, which includes: call activation methods, quiet calls, and calls for emergency crashes. (Image credits: Apple [1]).

Figure 2.6: Screenshot presenting the emergency call screen, which once activated will call the emergency services. (Image credits: Apple [1]).

Furthermore, many Android smartphones, including those made by Samsung, offer a range of emergency features that can be useful in the event of an emergency. Some of these features include the ability to call emergency services directly from the lock screen, a panic button that sends a distress signal to pre-selected contacts, and the ability to share your location with emergency contacts. Additionally, some phones come with a built-in medical information app that allows you to

store important medical information, such as allergies and current medications, which can be accessed by emergency responders in the event of an emergency. These features provide users with quick and easy access to emergency services and help ensure that they can get the help they need in the event of an emergency.

2.5.2.2 Mobile Applications for Emergencies

There are several mobile applications available for emergencies. Some examples include emergency medical information apps that allow users to store important medical information, such as allergies and current medications, which can be accessed by emergency responders in the event of an emergency. Other apps allow users to call emergency services directly from their phone, share their location with emergency contacts, or send a distress signal to preselected contacts in the event of an emergency [1]. These apps can be particularly useful for people with preexisting medical conditions, as well as for people who are travelling to unfamiliar places and want to have easy access to emergency services.

To give an example, the app "Life 360: for personal safety" Life 360 is an app that helps people monitor the safety and location of their loved ones [154]. It allows users to set up location tracking and receive notifications when people arrive or leave a location. In the event of an emergency, the app can be used to quickly locate friends and family and even make calls to them. The app also allows users to share their location and connect with friends and family through photos and text messages. Life 360 could be a useful tool for keeping track of loved ones and ensuring their safety, and it is available to both Apple and Android users [154].

2.6 Discussion

This chapter in addition to Appendix A has presented a state-of-the-art review for technology in crime prevention. Based on the resulting literature, this section will discuss this work and the wider academic and commercial context. A summary of the information discussed in this section is provided in Table 2.1.

Table 2.1: A classification of technology-driven solutions for the prevention of crime, describing the classification theme, the technical advantages, the limitations of the approach, and the potential application area of the technology.

Technology Classification	Technical Advantages	Limitations	Applications
Audio-Visual	Person identification Activity recognition Facial recognition Detecting street crime	Privacy – public capture of data Battery life – constant power needed Accuracy – no guarantee on data quality Technology misuse – potential for bad actors Equipment limitations – bad microphone, poor lighting, etc	Collection of evidence Predictive policing Detection of identified people height Deterring offenders Perception of safety
Ubiquitous Sensing	Tracking of known individuals Tracking of crowds Low energy use Social contexts	Limited individual functionality Technology misuse – potential bad actors Equipment limitations – MAC randomisation, RSSI issues, poor signal, etc	Stalker detection Context analysis Predictive policing Perception of safety Home safety
Data Science	Behaviour analysis Crime prediction Context analysis Violence detection Harmful Language detection	Ethics – permission of users Privacy – computational handling of data Data quality – data quality varies depending on hardware Data requirements – large amounts of data needed Affordability – cost is expensive for expertise/resources Data integrity	Detecting domestic abuse Detecting online abuse Detecting harmful content Deterring offenders Wide-scale analysis Understanding data Predictive policing
Information Technology	Easy to access dashboards Behaviour analysis Human-based analysis Visualisations	User-dependent Affordability – cost expensive Data requirements – accurate data is needed	Community policing Community reports Crime mapping Predictive policing

This review provides literature across a range of areas, first, audio-visual technologies were considered. For example, CCTV surveillance has been widely used to prevent and report crime in different circumstances. The continuous developments seen in computer vision allow for the on-board computation of different biometrics (i.e., face recognition or gait analysis) and for the extraction of patterns, which are key to identify suspects, as well as activities taking place within its field of view. Despite the great potential shown by CCTV surveillance to prevent various types of crime in public and outdoor places [46], several drawbacks are found in its application within home environments (e.g., abusive relationships), including privacy and typically visible placement. A home CCTV surveillance system would require the installation of numerous CCTV cameras, since the view field is limited to the room where the camera is mounted. CCTV cameras are sensitive to lighting conditions and occlusion (e.g., aggressive behaviour in the night could not be detected). Ultimately, as outlined in numerous works, there is a common reluctance to adopt video cameras in home settings due to privacy concerns [155]. Video-based technologies alongside the latest machine learning and pattern matching algorithms can be of great use to avoid crimes in outdoor settings, as well as in private businesses and public transport, but alternative sensing technologies are preferred for home environments.

Audio-based technologies have been used for many years as a crime prevention tool through the use of wire tapping and recording devices [156]. Furthermore, novel applications such as gunshot detection [157] and screaming / shouting detection [158] are surging with the advent of machine learning and pervasive computing. From a psychological point of view, compared to CCTV surveillance, audio monitoring is considered as a less invasive way of monitoring individuals [159]. Therefore, this indicates that privacy concerns found in the use of video surveillance systems in home environments are mitigated by the use of audio-sensing devices, making them an attractive mechanism for the collection of evidence. In addition, current advances in audio processing and machine learning allow the computation of various elements that can be crucial to prevent domestic violence. For example, consider a violent argument between a potential victim of domestic abuse and the abuser. Through audio diarisation, their voices can be separated, therefore, allowing for the analysis of each of the audio streams sep-

arately. Bad language within the conversation could be identified through the use of speech recognition. In addition, the mood (e.g., anxious, violent, sad) of each person could be estimated through the use of audio sentiment analysis techniques. Therefore, audio-based technologies can be of use as a means of obtaining evidence. Further advances in audio and NLP would grant this technology with crime prevention capabilities by identifying emergency situations inferred from sound.

Short-range communication technologies, such as WiFi and Bluetooth, have been shown to be useful in analysing the social context of a person [160, 161]. For instance, by continuously logging the presence of Bluetooth devices around a device and their respective RSSIs, it is possible to know the alone and accompanied periods of an individual. In addition, it is possible to infer the social context of a person by the analysis of the variation in the number of surrounding devices across time. For example, when a person is walking on the street, the social context could be inferred by the changes in the number of devices detected over time. Similarly, when a person is alone, the number of surrounding devices is not expected to vary. Furthermore, the closeness of an external device to a person could be evaluated by the long-term analysis of the corresponding logged MAC address of such external devices over time. This could be of great use to identify stalking scenarios and, more importantly, the periods of time when the potential stalker and the potential victim are found alone. However, an increasing number of devices make use of MAC randomisation, which can be a problem as it can make a system more unreliable in detecting local devices accurately.

Current portable devices such as smartphones or smart watches already incorporate Bluetooth transceivers, allowing for the continuous log of surrounding devices. In addition, current Bluetooth transceivers make use of a low energy protocol, BLE, reducing the power consumption required to operate. Another potential application with the use of Bluetooth device scanning is the prevention of robberies at home environments. To do so, Bluetooth Beacons could be installed at home environments to detect surrounding devices which remain within the Bluetooth antenna range for suspiciously long periods of time or/and at a suspicious time, raising an alarm when appropriate. Furthermore, short-range communication technologies could be employed for the monitoring of different

areas through the detection of intrusive scenarios. GPS technology can be of great use to ensure that security distances are met. Reported crimes in which the offender is forced to wear a GPS tracking device are a good example. However, despite the fact that most portable devices incorporate GPS technology, personal geo-locations are only shared with services the users opt to share the data with, therefore being not publicly available. Given this, GPS technology can be considered of little use in unreported cases of crime. On the contrary, short-range communication technologies constantly transmit a signal to surrounding devices, thus overcoming the drawback of GPS technology in this context.

The data science methods highlighted as part of this review indicate the widespread use of data science in crime prevention. For example, behaviour analysis, crime prediction, violence detection, and context analysis have all been considered in the data science section of this review. Data science enables a quick method of content moderation, such as social media websites to detect harmful content [38]. Similarly, data science methods are often linked to other methods identified within this review, such as CCTV video analysis [162] or crime mapping and predictions [91]. The reviewed performed in this chapter highlights the high level of use that data science methods have within crime prevention, due to easy to propagate methods that are easily reproducible. However, some notable issues in the literature exist, most notably a lack of data available in existing sources. Apart from text-based content, limited training datasets are available for machine learning algorithms, specifically when considering multimodal approaches. These multimodal approaches also indicate potential improvements that could be made to data science methods, in particular the potential fusion methods that could be used to take multiple modalities of data to assist in determining decisions. Based on the content of this review, this area is identified as having a high level of potential, due to the novel technologies and classifications which could be used by modern technologies.

Social media crime prevention systems make use of various data sources to predict the risk of a crime being committed. For example, as previous work shows [54], there is a common tendency for abusers to isolate victims of domestic abuse by controlling and monitoring their electronic devices and social media, with several reported cases of account hijacking or even destroyed devices. Given

this, information-based technology could be employed to identify deviations from the regular personal social media behavioural pattern of a potential victim of domestic abuse, therefore providing clues for the detection of further criminal actions. Similarly, when analysing the social media behaviour of individuals, it could be possible to employ probabilistic models to estimate the probability that a specific person will commit a crime.

Crime mapping techniques [163] that allow the identification of crime hot spots could be employed to optimise the use of personal and technological resources. Although the information gained from the aforementioned techniques could benefit the identification of potential criminal actions, various important limitations are found. First, the use of trend analysis techniques may require long-term studies, where the behaviour of individuals may need to be monitored for some time to establish a useful profile of activity. This can imply a long delay between the occurrence of the criminal actions and the conclusions drawn from the analysis. Other information technology techniques, such as mobile computing, are also identified. Mobile technologies and applications provide methods of reporting emergencies based on the current context of the users, whether this is in public or private spaces. The widespread use of mobile phones [36] has notably encouraged the development of applications and software to support personal safety. Most of these applications include methods for contacting emergency contacts, reporting crimes, or detecting emergencies based on on-board sensors. Mobile technologies are therefore still a novel area where devices and onboard sensors can be used to detect crime in near-real time. For example, sensor technologies in addition to processing power could enable more localised processing technologies through edge processing and lite machine learning models.

2.7 Challenges

This section presents the main research challenges faced in the development of computing technologies to prevent crime. A breakdown of these challenges as themes is presented in Figure 2.7.

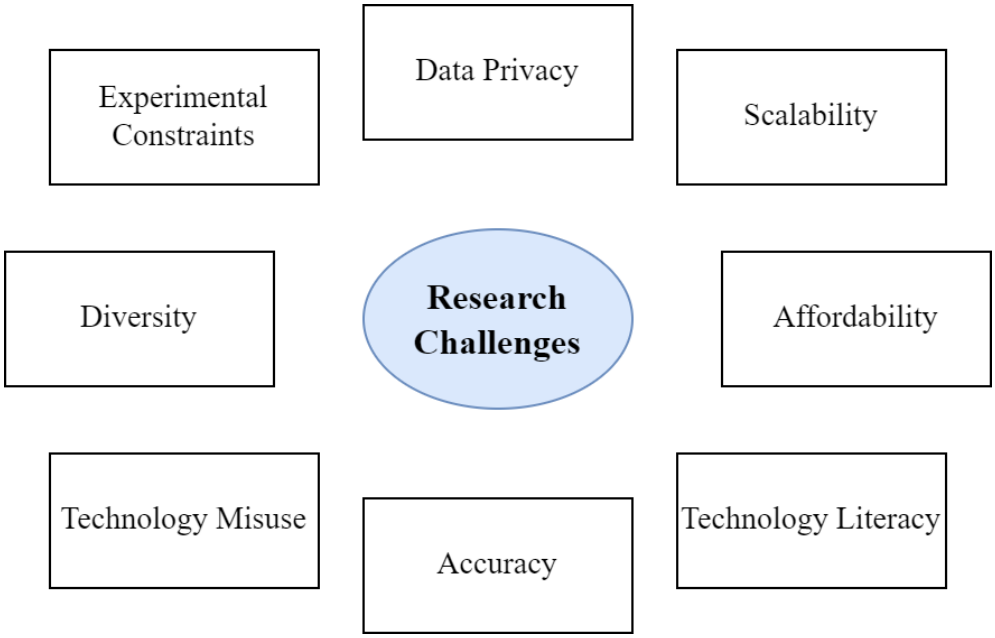

Figure 2.7: A diagram presenting the challenges of using digital technologies in crime prevention identified during the literature review. The challenges are presented as a collection of themes that researchers in crime prevention technology should consider.

2.7.1 Data Privacy

In recent years, research has been conducted on the privacy implications of smart devices. Although there is a huge privacy risk, as has also been stated in [164], system developers are normally required to meet strict deadlines to avoid losing competitive advantage, leading to immature products that do not meet the expected privacy and security requirements of their specific target applications [165]. A large number of users may feel understandably uncomfortable sending their personal data to remote data stores or clouds that they cannot see or control. Transferring user data over networks (including secure networks) makes systems vulnerable to theft and distortion. To be perceived as legitimate by the public and to meet ethical and legal standards, privacy requirements and concerns must be satisfactorily addressed.

The privacy concerns mentioned above can often be addressed by guaranteeing that user personal data will be processed at the point of collection and will not

be transferred remotely. By employing edge computing and on-device processing techniques, chipsets can potentially be exploited at or close to the data source to perform computational tasks instead of transmitting all of the data and processing them on a central server. A competent trade-off between personal privacy and computational power should certainly be optimised for each crime prevention tool if this is to be widely employed by the public.

2.7.2 Diversity and Scalability

Developing AI technologies to detect explicit personal behaviours or actions requires the collection of data from a large number of users, to enable predictive models to incorporate sufficient variability in order to generalise well on unseen data. Therefore, it is crucial to carry out data collection processes on a large number of experimental participants, as well as to consider the differences that may exist between different groups of individuals. For instance, if someone is planning to develop and launch a novel audio-based system to recognise violent behaviours for the prevention of domestic abuse in households within the UK. The first problem that a system developer may face is that not everyone in the household speaks English. A speech recognition model that is only trained with data from native speakers would certainly not perform well with data processed from nonnative speakers. Additionally, there are other sound characteristics, such as tone, phonetics, intonation, and melody, that may vary considerably between different languages or even between different accents of the same language. Thus, although it may involve tedious processes of data collection, it is crucial to identify the target group and incorporate an adequate level of inter- and intra-group variability when developing intelligent systems. This issue becomes even more crucial when dealing with sensitive matters like crime where false negatives can translate into serious personal health-related consequences.

2.7.3 Accuracy and Experimental Constraints

Intuitive optimisation of various parameters often results in better and more accurate models, with continuously lowered error rates. However, there exist various limitations and challenges that do not normally allow AI applications

to exhibit error-free performance. First, machine learning and deep learning algorithms typically require large amounts of data to successfully undergo the training phase.

In addition, such data should incorporate adequate inter- and intra-subject variability for the classification or regression models to be able to generalise well on unseen data samples. This means that in addition to the need to collect large amounts of data, those data have to be variable enough to adequately represent the characteristics of the target population and avoid large generalisations or out-of-sample errors. In addition to ensuring the collection of "enough" data samples, adequate feature engineering and machine learning techniques can be a crucial factor in optimisation of the accuracy achieved by AI-based crime prevention systems. Therefore, research challenges arise from the need to improve the performance achieved by the state-of-the-art in the corresponding fields.

For example, as a current survey on sentiment analysis indicates [166], the classification precision achieved by the state-of-the-art on sentiment analysis using audio recordings is in the range of 72.9% to 85.1%. This means that if an audio-based verbal abuse system were developed, it would be wrong in 14.9% of the cases. The behavioural data required for models [167] such as video processing [168], emotion recognition [155, 169], and activity recognition, must be collected in highly variant free-living scenarios rather than in controlled settings.

2.7.4 Affordability

Cost is a key factor in implementing crime prevention technologies, as end users and other supporting organisations would need the device to be as inexpensive as possible if it were to be adopted into practise. In this regard, while tagging technologies are relatively inexpensive, applications that require high processing power may depend on expensive hardware.

Edge computing enables processing of all or part of the data at the location where it is collected. Data that are only ephemeral important can be crunched on the edge device itself. This is in contrast to cloud-based systems, where data is sent to large, remote data centres for processing. According to Moore's law [170], small devices at the edge have become more computationally powerful. If the

trend continues, it is only inevitable that switching to edge platforms would offer much more affordable solutions in the long run. This territorial proximity to the endpoint is good for both latency and efficiency, as it saves networks from unnecessary congestion and from the carriage of sensitive personal data.

The ongoing minimisation and mass production of electronics has enabled a reduction in the cost of edge computing devices, whereby various complex computations (such as AI and data processing) can take place on-board. However, the more data and computing intensive an application is, the more data storage and processing power is required, with this having an impact on the price to be paid by end users. In this regard, the adoption of high processing power technologies by the public has two main research challenges associated with it. First, to keep up with the ongoing miniaturisation and cost reduction of electronic components. And secondly, the optimisation of signal processing and machine learning algorithms so that these can be adopted using a lower computing power.

2.7.5 Technology Misuse and Literacy

Advancement in technologies enables law enforcement and voluntary services to support victims and reduce or prevent crimes. However, the increasing complexity and communication capabilities present in these technologies have also opened new pathways for data interference [171].

Little empirical research has been published on the use of technology in stalking of intimate partners, as most current efforts focus on online abuse on social networks or texting [172]. Within these, the work in [172] conducted a survey with 152 advocates for domestic violence and 46 victims. The study found that modern technologies can potentially give perpetrators multiple tools to control and manipulate people and that technology-facilitated stalking must be treated as a serious offence. Therefore, there is a need for non-judgemental responses from service providers and law enforcement to victims experiencing such abuse. As practitioners observed, advising victims to turn off devices, remove social networks, or change profiles or telephone numbers puts an enormous burden on the victim to adjust their behaviour [173]. Additionally, the disengagement of technology can mean that victims are increasingly uncontactable, which can affect

the type and timing of the support they receive from services.

There is a great need to increase public awareness of the use of spyware to commit abuse and stalking. Similarly, law enforcement and victim support and rehabilitation organisations will benefit from learning about the latest developments in technologies that could vulnerable individuals or prevent technology misuse.

2.8 Opportunities

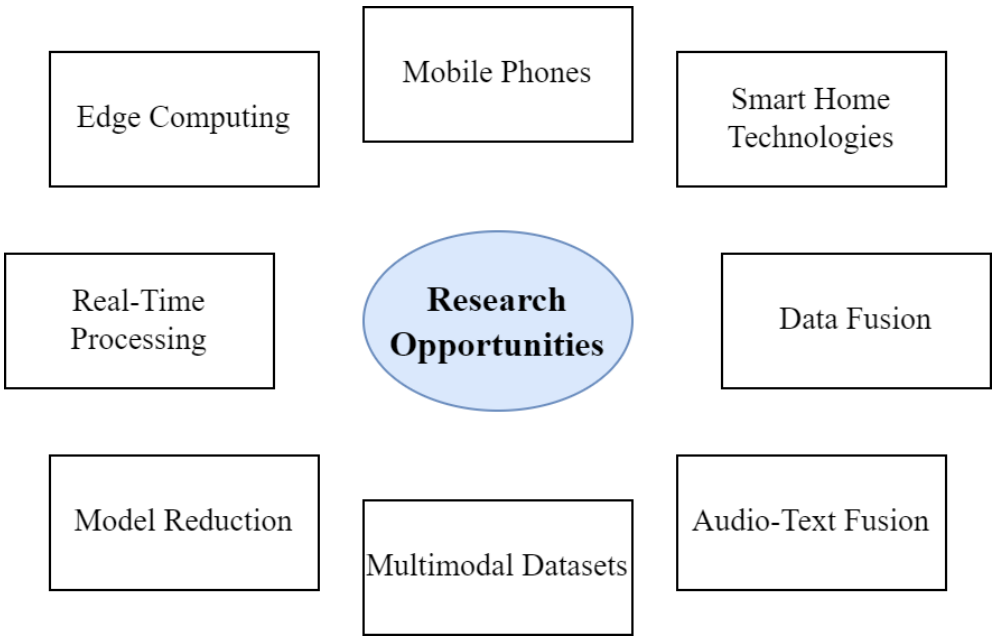

Figure 2.8: A diagram presenting the opportunities of using digital technologies in crime prevention identified during the literature review. The opportunities are presented as a collection of themes that researchers in crime prevention technology should consider.

This section presents the opportunities identified for the use of technology for crime prevention. A breakdown of these opportunities as themes is presented in Figure 2.8.

2.8.1 Edge Computing

Traditionally, most of the processing for data-intensive applications was done on a central cloud to take advantage of the fast and powerful computing infrastructure. However, major issues related to latency, security, and privacy can be identified in the use of cloud-based systems for crime prevention. A new trend is emerging to process data on the edge. The motivation behind edge computing is to perform data processing as near as possible to the point of data production. With this, privacy and latency issues present in cloud-based systems can be significantly mitigated. AI is gradually finding its way into embedded systems that are becoming smaller and less power demanding, while offering fast processing power and low latency at an increasingly attractive cost.

Several standard edge computing devices suitable for carrying out heavy signal processing and machine learning applications are already available. For instance, both Nvidia and Google have recently released their respective development boards, namely Jetson Nano and Google Edge TPU, with the aim of enabling users to develop and run AI applications on the edge. In addition to their portability and the privacy advantages they offer, such boards are supported by sophisticated development kits that consist of a system-on-module connected to a development board that incorporate numerous connectors like USB and Ethernet to share the data gathered when desired. Furthermore, the above devices also support major deep learning frameworks and tools such as TensorFlow. In general, these observations, along with current developments in the field [174], suggest that edge computing is finding its way into city centres for early detection and prevention of criminal actions.

2.8.2 Smart Homes

Smart home devices are devices that can be remotely controlled with smartphones and voice commands, and are designed to be used in the home environment. These devices include appliances, lights, thermostats, security cameras, and smart speakers. They allow users to remotely control and monitor home appliances, lights, and temperature, set schedules, and create scenes that control multiple devices with a single command. Smart home devices can be used to make a

users home more energy efficient, secure, and convenient. Popular examples of smart home devices include smart locks, smart lights, smart doorbells, and smart security cameras. Smart home technologies provide a potentially novel method of using technology to prevent crime, notably smart home devices could be used to monitor strange behaviours or identify home use while a user is on holiday. The domestic nature of smart home devices also provides an opportunity to prevent crimes that happen within the home, however, the privacy aspect of this approach should be considered in future work.

Smart home devices are becoming increasingly popular, cheaper, and accessible to more people. They can also be integrated into other smart devices to create a single-app controlled ecosystem, and they can also be integrated into virtual assistants such as Amazon Alexa, Google Assistant, and Apple Siri, making them more easily controlled. Virtual assistants and smart speakers provide another method of sensors in the domestic environment, which also raises opportunities for future implementation and privacy concerns for the user's home environment.

2.8.3 Data Fusion

Data fusion in machine learning refers to the process of combining multiple sources of data to improve the accuracy and performance of a machine learning model. This can be useful when working with complex and heterogeneous datasets, where different data sources may provide complementary information that can be used to improve the model's predictions. There are several different approaches to data fusion in machine learning, including:

1. Feature-level fusion: This involves combining different data sources at the level of individual features or variables, such as by concatenating or averaging the values of multiple features.

2. Decision-Level Fusion: This involves combining the predictions of multiple machine learning models, such as using ensemble methods or voting schemes.

3. Deep learning-based fusion: This involves using deep learning techniques, such as NNs, to combine multiple data sources and make predictions.

Data fusion can be useful for a variety of applications, including image recognition, NLP, and predictive modelling. Using information from multiple data sources can help improve the accuracy and performance of machine learning models and enable them to make more informed and reliable predictions. For example, Crime Telescope [175] is a website that collects data on web crime, urban data, and social media to predict hot spots in crime. Using statistical and linguistic analysis, key characteristics are extracted and crime hotspots identified. It also offers visual representations of hot spots on interactive maps. The study showed that combining different types of data increased the accuracy of crime hotspot predictions by 5.2% compared to traditional methods. The user-friendliness of the platform was evaluated through surveys.

Another example of data fusion in crime prevention is an article [176] that proposes methods for improving crime prediction using environmental context information using a deep neural network (DNN) that combines data from various sources such as crime statistics, demographics, and meteorological data. The model is trained using Chicago's crime-related data, which indicates that it is more accurate than other prediction models. A pedestrian detection algorithm has also been proposed [177], where they use multisource face images and face recognition algorithms based on mixed orthonormalised partial square regression analysis for accurate results under complex lighting conditions. The algorithm combines statistical learning theory with combined orthonormal partial square regression and modified SVMs to improve performance in small sample scenarios. The experimental results show that the proposed algorithms outperform other state-of-the-art methods.

2.8.4 Real-time Processing

Real-time processing is the system's ability to process and react to data and events in real time without significant delay. In the field of video surveillance, real-time processing would analyse video recordings, quickly detect potential crimes or accidents, and respond by sending immediate information or alerting authorities. On the contrary, this is not a system that analyses the video only after it has been recorded, which can cause a delay between the event and its detection.

The existing literature [88] is focused on the use of video surveillance systems to detect and prevent crime. Current systems are often passive and are used to gather evidence after the crime occurred, but the literature proposes that "intelligent video surveillance systems" be developed using deep learning technology to actively monitor the crime in real time and quickly detect and respond to it through real-time notifications and multimedia on the Web. Another paper [178] that looked at crime prevention tools using real-time processing discusses the growing use of crime maps as a crime prevention tool. They pointed out that until recently, few criminal justice authorities were able to create crime maps, and few investigators were able to examine crime spatial distribution. However, due to advances in mapping technologies and crime prevention theory, the interest of scholars and practitioners has increased significantly. The paper also discusses the potential challenges and pitfalls of real-time crime mapping and the possible progress to be made in the coming decades. Real-time processing could therefore enable crimes to be detected as they happen, or provide a much quicker Police response to a potential situation.

2.9 Conclusions and Directions for Future Work

Different technology-driven solutions and the potential adoption of contemporary smart ubiquitous, machine intelligence systems and miniature technologies for crime prevention have been considered and evaluated. There is enormous potential associated with the adoption of short-range communication technologies as a tool for crime prevention. These technologies can be employed alongside current approaches, such as GPS monitoring, to provide a more robust proximity detection system that can detect proximity in both indoor and outdoor environments.

Furthermore, the use of audio-visual technologies and user devices provides significant opportunity due to widespread adoption for future research in crime prevention in both outdoor and indoor areas. Violence detection scenarios in home or work environments and the use of audio-based technologies appear to be preferred against that of CCTV cameras given the ubiquity, spherical field, and reduced privacy concerns exhibited by the former systems. As exposed, the

use of audio signal processing along with machine learning techniques, allowing applications such as speaker diarisation, speech recognition, person identification, and sentiment analysis, could be key to identify violent language as well as violent actions. From a technical point of view, standalone systems employing single-sensing technology exhibit distinct limitations. However, the combination of technologies can be of great use in identifying violent scenarios, which can lead to the prevention of further occurrences and therefore to the ultimate prevention of criminal activity.

AI also provides novel opportunity based on the reviewed literature, with technologies such as NLP and social media analysis becoming critical to preventing harmful content and online crime. Should these technologies be further adapted to smart homes or edge devices, opportunities for preventing non-digital crimes using digital technologies would increase. These approaches could enable smart homes to be safer for victims by analysing audio and textual content from a conversation, while also increasing privacy confidence through edge-based processing. Limitations were identified in this review, such as the lack of existing datasets for this approach and the complexities in working with fusion data on the edge.

In conclusion, not all circumstances or situations are the same, and there will be no 'one size fits all' solution. As such, it is important to explore heterogeneity in offenders, victims, offending contexts, and offending patterns to understand which solutions might work for whom and under what circumstances. Besides privacy, scalability, affordability, miniaturisation, and personalisation are some of the important factors that need to be considered when designing technologies for crime prevention. Nevertheless, future work should focus on exploring the use of crime prevention technologies and the development of novel solutions. The work should look to explore how more complex data processing, such as data fusion, can be performed by user devices using some of the technologies identified in this review, while also considering the privacy and accuracy of the data collected.

Chapter 3

Research Design

3.1 Chapter Overview

The multidimensional nature of speech and conversation presents a demanding task for computational machines to understand how individuals communicate with one another. Understanding a conversation is highly subjective to the individuals involved, due to the merger of speech, communication patterns, and the surrounding context. Individuals take on different personalities by different triggers, and therefore there is a need to employ multiple modalities simultaneously. This challenging environment becomes even more complex when violent or harmful language is incorporated, as inaccurate detection could lead to severe consequences. Ultimately, violent acts consist of individual human beings inflicting harm on other human beings. The emotional states of the perpetrators at those moments can have a decisive effect on the degree of violence and even on whether or not aggression occurs at all. On the other hand, aggression is a behaviour, emotion is a feeling state, and therefore the links between aggression and behaviour involve relationships between objective actions and subjective feelings [179]. Several studies reveal that intimate-partner relationships can provide a potential context for intense emotions and conflicts, which can result in serious injuries or even death [180].

Given the ambiguity and versatility of personal emotions, their accurate recognition becomes an extremely challenging task that has been investigated by sev-

eral researchers [181, 182]. Researchers have attempted to detect and recognise emotions using speech recognition [181, 183, 184], image analysis [85, 185, 186] and NLP which is often applied to expansive datasets formed through the use of social media services such as Facebook, Twitter, and YouTube to understand the emotion of the interaction [187–190].

The objective of NLP is to interpret, decipher, and make sense of human languages in a manner that is valuable. Therefore, emotion analysis, including negative emotions and behaviour analysis, could help detect the level of potential violence in human interaction. Recently, there have been multiple research efforts that focused on detecting negative emotions such as abusive language [191] and bullying [55] in online conversations; however, it is much more challenging to obtain a recorded conversation that is happening behind closed doors in real-life settings. Focussing entirely on text means that the analysis misses the vocal cues and other sound clues available in the environment when the conversation occurs. Violence characterisation is subjective in nature, which makes violent incidents usually manifested through characteristic audio signals (e.g., screams, gunshots, etc.). Sound (including low-frequency sounds) is often used in movies to elevate tension. Audio signals for violence detection are often used as an additional feature when abrupt changes in the energy level of the audio signal are detected using the energy entropy criterion [192].

The field of multimodal violence analysis in conversations has not received much attention. Inspired by these observations, this thesis explores through the power of technology and the ability of momentary assessment of real-world data, if it is possible now to go beyond sentimental analysis, hate, and abusive language detection online to detect violence during verbal conversations. Therefore, the methodology presented in this chapter explains the design science research methodology being used and how the artifacts designed through this method relate to the overall aim of the work. The data of the project through the dataset curation and processing is explored, before the computational equipment and ethical considerations are presented.

3.2 Methodology

The project is industry related through its relevance to policing and domestic violence, and therefore it is necessary for the research methodology used to ensure that an effective process is used throughout. For this purpose, the design science methodology was used. Design science focusses on the development and performance of designed research artifacts, where artifacts are developed contributions or outputs of the study. Design science, in contrast to natural sciences, considers the potential of research applications and relates to devising artifacts to attain research goals [193], the methodology is common in novel subject areas such as human-computer interaction (HCI) and algorithms where information technology is used [194]. In its basic form, design science is about understanding and improving the investigation of potential research projects to develop artifacts that intend to solve a problem [195]. Therefore, the design science methodology is relevant to the research conducted as part of this thesis, through the development of research artifacts to detect violent conversations on the edge in the domestic violence context.

In information systems design science research, some methodologies identify six steps that include problem identification and motivation, definition of the objective for a solution, design and development, demonstration, evaluation, and communication [194]. Therefore, each of these steps has been completed as part of this research project. Problem identification and motivation has been identified in the introduction in Chapter 1 and from the literature reviewed in Chapter 2. The definition of the objective for the solution has also been identified in the introductory Chapter 1 of this thesis. The design and development of research artifacts is the focus of chapters 4, 5 and 6 where specific research design, development and analysis tasks are undertaken. Demonstration and evaluation of the research artifacts is presented in Chapter 7 where discussion and evaluation is provided of the overall solution. Finally, throughout the project, communication has occurred with industry and other researchers, satisfying the design science research communication process identified through organisational workshops.

3.2.1 Research Artifacts

The design science research methodology investigates novel problems through the development of research artifacts [193]. Due to the nature of the study undertaken throughout this thesis, multiple artifacts are developed, each main thesis chapter relating to a separate element of artifact development. For example, this methodology chapter presents a novel multimodal dataset for linked audio-text content from television shows, which has a wider contribution to both science and industry due to pre-trained models based around UK accents and television content. This demonstration of the use of artifacts in the research highlights how artifacts guide the development process. In addition, research artifacts enable rapid communication of results through conference proceedings and feedback meetings with industry partners. The use of research artifacts means the results of the research can be rapidly demonstrated and propagated through industry and research, leading to improved levels of feedback from potential impact areas. The artifacts used as part of this research also build upon each other, meaning that early audio fusion approaches are still included in the final edge device, highlighting the ongoing investigations that occur.

3.2.2 Study Population

Domestic violence is a widespread problem that affects millions of people around the world and the consequences can be catastrophic for victims and their families. The target study population for this study is victims of domestic violence or abuse seeking help and protection from authorities such as courts and police. This includes all types of vulnerable people, such as elderly citizens in care homes; foster kids in the fostering system with a troubled past, and for abusive households. A crime prevention tool for people in such environments can help them get justice and the help they need, as it would enable them to gather evidence through a legal system that would all be tracked and recorded in official documents. However, due to the sensitive nature of the study and the early stage of the work, it was determined that engaging directly with at-risk individuals was not appropriate or necessary and therefore, for the purpose of the investigation, a custom dataset was curated.

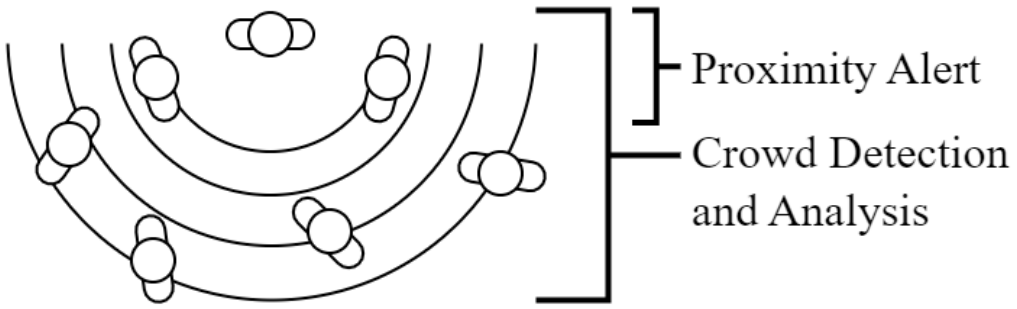

Figure 3.1: Illustration demonstrating how a system developed using wireless sensing could be used to identify crowds to perform analysis. Such a technology could also be used to identify proximity, especially in the case of Bluetooth Low Energy.

3.2.3 Organisation and Practitioner Workshops

A series of workshops was conducted from June 2018 to May 2020, gathering academic experts, practitioners, engineers, technologists, and representatives from diverse sectors such as Police forces, government agencies, charities, trusts, voluntary support groups, and end users. The workshops aimed to explore the opportunities offered by emerging technologies for the monitoring and prevention of domestic abuse. Two additional workshops were dedicated to sharing the research results with Nottinghamshire Police and Met Police. The following is a condensed summary of the main insights derived from the initial workshops:

1. Exploratory Workshops Two workshops were organised to explore the role of technology in reducing the risk of domestic abuse. The primary objective of the first workshop was to collaborate with experts in the field and create a detailed "requirements brief." This brief outlined the priority issues related to domestic abuse and identified the current limitations and potential future approaches to address them. In the second workshop, a diverse group of participants, including domain experts, designers, and engineers, came together to brainstorm and propose potential solutions to the identified problems.

2. Co-Design Workshop: The workshop promoted discussions of existing crime prevention technologies, delving into their potential. The attendees also ex-

amined the challenges faced by survivors when seeking help or gathering discreet evidence. The session introduced several technologies that could help victims, aid collecting evidence, and enhance law enforcement efforts. In particular, participants showed a keen interest in the portability of compact edge computing platforms, such as wearables, valuing their distinct interaction methods compared to smartphones.

The prospect of proximity detection techniques, capable of identifying individuals and detecting abnormal activities in the surrounding environment, sparked enthusiasm among participants. This feature resonated particularly well with the struggles faced by vulnerable individuals in documenting such incidents. Moreover, the idea of IoT devices providing evidence and serving as reliable journals intrigued attendees, as these technologies were novel to them.

3. Co-Creation Workshop: During the workshop, a focused discussion was held on potential technological solutions, including proximity detection and tagging, accompanied by practical demonstrations. One idea explored the use of short-range tagging technology as part of a panic alarm system. This involved placing small tags or "dots" in homes or public spaces, serving as triggers for alarms or access points for information. Furthermore, proximity detection technologies [196, 197] were examined for their ability to enforce the legal distance between potential offenders and victims, as shown in Figure 3.1.

   Another concept involved using devices connected to the Internet, such as smart speakers, to automatically identify instances of abuse as they occur. To reduce false alarms, the system could analyse multiple indicators, such as door slam, raised voices, specific keywords in conversations, or other signs of potential threats or risks. Practical aspects were considered, including maintaining user privacy and customising trigger actions to individuals to enhance system sensitivity. When triggered, the data could be stored locally or in the cloud, with alerts sent to trusted parties for further action.

4. In-depth consultations were conducted with end-user organisations, build-

ing on the insights gained from the previous workshops. The primary objective of these online discussions was to narrow down the potential development paths and to create a series of prototypes on paper. Each prototype was carefully reviewed and discussed with the respective organisation to gain a deeper understanding of its implications and limitations. This process aimed to ensure a thorough exploration of the practical considerations and challenges associated with implementing prototypes in real-world contexts.

5. During the Design Evaluation Workshop, further refinement to the proposed prototypes based on the earlier discussions on various technologies. These prototypes included the use of short-range communication methods to detect and document suspicious activities in close proximity, as well as the integration of artificial intelligence to identify violent behaviour. The workshop participants expressed a strong preference for a single-edge device dedicated to crime prevention and evidence collection while prioritising user privacy. Using edge computing, data processing could be performed locally, eliminating the need for remote storage. The substantial processing power of the device would enable complex tasks like voice and sentiment analysis to be conducted directly on the device. This approach would allow the development of customised and adaptable interfaces or functionalities that cater to the diverse needs of different user groups.

The series of co-design and co-creation workshops revealed a clear need for innovative technological solutions in the field of crime prevention. It became evident that a generic, one-size-fits-all approach would not adequately address the diverse range of issues and user perspectives. Instead, a collaborative effort involving multiple organisations is essential to develop optimal solutions that cater to specific contexts and user needs. The workshops emphasised the importance of collective participation in tackling the complex challenges associated with crime prevention, highlighting the need for tailored and comprehensive approaches.

3.3 Data Collection

In machine learning, data collection and labelling is the process of adding one or more meaningful and informative labels to preidentified data to provide context so that a machine learning model can learn from it. The common approach in NLP to engineer labels is the detection of keyword-based hate speech using ready-made lexicons [124,198]. Since there are no common or agreed lexicons for violent words, the data was collected and labelled by three reviewers to suit the needs of the study. Although utilising lexicons such as the HateBase, the results of the systems do perform high, it was decided not to be used as they can be highly biased and unreliable. Furthermore, it is challenging to maintain and maintain the lists up-to-date [199]. Waseem et al. [200] study was applied as a guide, as the authors made a generic definition based on hate-related content found on social media to address the problem of detecting it on Twitter. Gender Studies and Critical Race Theory (CRT) are used as a baseline. For the study, they attempted to annotate a total of 16,849 tweet corpus into three categories: 'Racism, Sexism, and None'. To ensure that the corpus was reliable and impartial, they had a "a 25 years old woman studying gender studies and a non-activist feminist to reduce annotator bias" [48].

The lack of existing datasets presented a challenge in relation to undertaking the design science methodology, it was necessary for the research artifacts to use an example data source that can provide a demonstration of the research contribution. Most existing datasets are based on existing corpus of text-based data such as tweets, or video content that does not match a real-world setting. For this research, a dataset that linked both text-based transcriptions and audio segments was needed, with each of these segments relating to an audio file path and a transcript field in the dataset. Therefore, it was necessary for the data curation, labelling, and processing activities to occur prior to work on the machine learning models and fusion techniques.

3.3.1 Dataset Curation

Based on the lack of existing datasets, it was necessary for a new dataset to be curatred for the implementation of the system. For this purpose, a number of

relevant videos from television series and other audiovisual material containing domestic abuse scenes with violent and abusive conversations are retrieved and the corresponding audio signals are extracted from the videos. A posteriori, a total of 1295 audio files are divided into time segments with a length of 10 seconds and labelled according to whether they contain verbal abuse (including violence) or not on a scale from 1 to 5. They are derived from incongruous British television series and the different elements of the proposed dataset are described as follows:

1. A collection of 50 minutes and 42 seconds of video data from scenes from the series Eastenders [201]. With this, 304 segments of audio with a duration of 10 seconds are obtained, 195 without verbal abuse, and the remaining 110 segments with verbal abuse. Another 133 10 second segments were retrieved from 4 to 7-minute long YouTube videos. Within these, 73 were nonviolent and leaving 60 with violence.

2. A set of 507 audio files derived from 5 to 10 minute videos from the British TV series Coronation Street [201] were collected and split into 10-second segments of audio files. From these, 352 do not contain any verbal abuse leaving 288 segments with verbal abuse.

3. The corpus also includes scenes that showcase violence and abuse from another popular series named Emmerdale [201]. From this series there are a total of 350 10-second audio clips from which 115 did not contain any violent language leaving 235 with abusive language.

Therefore, the resultant dataset embodies a total of 1295 segments, of which 633 contain verbal abuse and the remaining 662 do not contain any form of verbal abuse. Implementation was primarily carried out on Google Colab as it provides a single 12GB NVIDIA Tesla K80 GPU that can be used for up to 12 hours continuously. The Python library ibmwatson was used to process the video files into 10 second segmented audio files.

This is the reason for limiting the size of each segment to 10 seconds to provide consistent functionalities across the audio and text features. While longer segments contain more information, violent and non-violent events will be mixed. To convert audio to text, two approaches were used to eliminate any errors. The

first approach was 'SpeechToTextV1' which transcribed the audio files into text; however, due to the British accent, the transcription was sometimes incorrect and had to be checked by the data reviewers as they labelled the text. It was not possible to obtain recordings of real-life conversations given the personal nature and the sensitivity of the content.

3.3.1.1 Data Labelling and Annotation

For this work, three data reviewers were recruited to annotate each transcribed speech segment; one of them being a graduate in Sociology who had a focus on domestic abuse and currently works with students from troubled homes, another who graduated in Forensics and Digital Security whose research focus was on violent language detection, and a final reviewer who studied Law whose expertise is in Criminal Law. Each reviewer assigned a score to each segment within the dataset with a violence intensity between 0-5 on a Likert scale. This intensity was accompanied by notes on the scale for each reviewer, who was unable to see any previous review attempts. Annotator disagreement was resolved by averaging the labelling scores, with these averages then becoming the resulting values of the dataset.

A posteriori, the resultant segments' labels were obtained by averaging the scores given by the three labellers as the sentiment polarity. The three labelers were used to mitigate natural subjectivity and complex feelings of language, and define mean values to be used as a curated set of resultant scores. To test the reliability of the scores, a one-way analysis of variance (ANOVA) test was performed and no significant differences were found in the mean rankings provided ($p = 0.932 > 0.001$). The mean ranking for each reviewer was 1.21, 1.23 and 1.22 with standard deviations as those presented in Figure 3.2.

3.3.2 Processing

The collection and curation of the dataset required processing to be performed, to label and transcribe the audio segments, in addition to providing the data in the required format. The novel nature of the dataset collected as part of this study makes it necessary to report on this processing, enabling future iterations of the

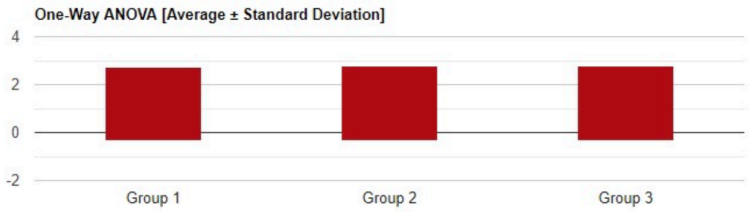

Figure 3.2: Average and Standard Deviation values of the three different reviewers rankings presented as a bar chart in which each group represents a separate reviewers responses. Limited variation was identified from the rankings indicating similar values being selected throughout this process.

dataset to be collected. Two processing scenarios occurred, with a significant pre-processing result, followed by a later post-processing result for the data to be used in the models presented in this research.

Table 3.1: A demonstration extract of the dataset produced as part of the research project including the linked ID field, the transcript, and the reviewer results. Values R1, R2, and R3 represent the different reviewer rankings and L represents the final determined label.

ID	Transcript	R1	R2	R3	L
eeaudio030	“are you suggesting i don’t go? well this is gonna make you feel uncomfortable uh i don’t know i’ve got people coming in for a few days at home will do good ”	1	2	1	2
eeaudio001	“whatever hey hey hey hey no tip allows you to do that my friend take your enough for data take your hands off my husband ”	3	2	3	3
eeaudio061	“something else wouldnt it? wouldnt it? Cause there’s always something else isn’t it. I cant live like this. No....no Chantelle please.... look i just ”	2	3	3	3
eeaudio024	“favorite wine don’t start another fight please don’t act the victim ”	4	4	4	4
eeaudio222	“come on come here what’s the matter? well i thought you’ll be angry. how could i be angry? ”	0	0	0	0
eeaudio232	“do not tell them about my job!I won’t. This is all your dad’s ever wanted me made a fool. yeah what are you doing distracting hubby from his work? oh ”	4	4	4	4
eeaudio248	“do not tell them about my job!I won’t. This is all your dad’s ever wanted me made a fool. yeah what are you doing distracting hubby from his work? oh ”	4	4	4	4
CS2050	“We can’t go sacking people for no good reason. I have got 16 stone of a reason apart feeling. We can’t fire people cause you don’t like their husband. Anyway listen she takes us to an ”	0	0	0	0
E2046	“You are not here to understand.Well why am I here then you haven’t asked for any ransom not that I know of anyways. They couldn’t pay even if you had. Look shut up or the guys is going back on.”	2	2	2	2
E2050	“One false move and the spikes will get you.”	2	2	3	2

3.3.2.1 Pre-Processing

Pre-processing of the dataset was performed using a custom Python script; the research required 10-second segments to be used and for audio segments to be linked with transcription segments. Data was pre-processed using the Python library FFmpeg, which enabled the videos to be converted to audio, and the audio segments to be converted into 10 second-segments. The reviewers could then listen to each 10-second segment and complete the relevant fields, making the process easy to identify and limiting any potential input errors from the data reviewers. Upon completion of the automated and manual transcriptions, data was placed into a CSV file as demonstrated in Table 3.1. Therefore, each individual row had details related to the ID (and therefore the filename) of the audio segment, the transcription information, the final review scores of each reviewer, and the final output score on the Likert scale of 0-5. It was also necessary for any resulting data to continue to link to a folder output location, and retain details of the original video file.

3.3.2.2 Post-Processing

Post-processing of the dataset was performed to ensure that the data fit with the required audio-text modalities. For this, only a limited amount of processing occurred, mainly related to converting the dataset to a binary classification format. The resulting post-processing result retains the results of Table 3.1 and converts them to binary classification, keeping label 0 as 0 and labels 1-5 as 1. This resulting dataset was stored in both CSV and JSON formats to enable the models to be trained on the data source. This post-processing process was checked by examining the data for errors, during which the reviewers did not make changes to the violence classifications determined. Processing for individual models is described in the relevant individual chapters.

3.4 Computational Equipment

The equipment used for the study was mainly processed on a 2.5 GHz Dual-Core Intel Core i7 Macbook Pro 2017, which enabled Linux-based commands to be

used. Most Python processing, which includes model training and running, was performed on Google Colab as it provides a single 12GB NVIDIA Tesla K80 GPU which can be used for up to 12 hours at a time. Google Colab also allowed the code to be reproducible and rapidly demonstrated, due to the online nature of the processing and the potential for extra context to be described. The specific equipment used for each element of the study is discussed in the relevant chapter.

3.5 Ethical Considerations

Ethical concerns were considered throughout the study, most notably due to the sensitive topic area and the likely sensitive data that would be included. It is due to the sensitive nature of these data that no individual interviews and studies were conducted. Instead, the study focused on publicly accessible data, which, while not entirely realistic, provided a reasonable evidence-backed approach to completing the required work. For example, using YouTube videos of UK-based drama series enabled a dataset to be rapidly produced and demonstrated, which allowed public sharing of the data more than any individual conversations. In addition to this, concerns related to the sensitive nature of conversations would make it difficult to find individual conversations that could be used in the models. Despite this, the research was continuously checked against known published literature and experts to ensure that the assumptions being made were realistic and based on expert knowledge, ensuring that while sensitive data were not collected, it could be potentially included in future iterations of the work.

Ethics was also considered in relation to the nature of the study; it was initially proposed that conversations could be recorded from actual participants; however, the private nature of participants, the ethical requirements, the documentation needed, and the sensitive nature of discussions meant that this was also not possible while testing the system. Instead, it was determined to ensure that the research was computationally focused, enabling future work to explore how such a system could be ethically tested in the future. These ethical considerations meant that the work was allowed to be published and demonstrated, following the design science methodology, and also allowed the work to be iterated with future knowledge and data collection. These ethical considerations throughout

meant that no sensitive data was collected or processed during the study, and therefore no sensitive data is published in this thesis or related documentation.

3.6 Conclusion

This section provided an overview of the methodology used as part of the research project. A design science methodology was used due to the process of developing and designing research artefacts, these artefacts are then presented as demonstrations within the remainder of this thesis. Additionally, details of expert interviews, workshops, and feedback are presented which provide an industry and policing point of view to the work produced. The section then provides an explanation and demonstration of the dataset generated as part of the work. A novel audio-text dataset is demonstrated, and details of the data dimensions and classifications are provided. The data has potential for future use, due to being the first of its kind known to the researchers working on this project. This methodology chapter provides an overview and context for future chapters in the thesis, and describes the methods used to ensure scientific rigour throughout the research.

Chapter 4

Natural Language Processing for Extracted Speech

4.1 Chapter Overview

This chapter investigates the application of NLP techniques to prevent crime. We begin by providing an overview and introduction to NLP, highlighting its significance as a powerful tool in various scenarios. NLP is a subfield of AI that focuses on the interaction between computers and human language. It enables computers to understand, interpret, and generate human language, allowing for a wide range of applications such as language translation, sentiment analysis, text summarisation, and speech recognition.

The chapter begins with an introduction to NLP and its significance in contextual scenarios. The focus of the overall chapter is on detecting abusive language from audio transcriptions. The chapter provides a comprehensive analysis of NLP techniques employed in classifying spoken language as abusive or non-abusive. It discusses models such as CNNs, RNNs (including long short-term memory (LSTM)), and BERT, showcasing their effectiveness in abusive language detection. The chapter concludes with a discussion of ongoing research to advance NLP techniques to detect violent language. By using NLP, it is possible to create safer environments and combat abusive behaviour in different contexts.

4.2 Background

The proliferation of natural language writings in the connected world makes it difficult to disseminate information and wisdom in a timely manner. The amount of information accessible makes it more challenging for people to manually process and thoroughly analyse text. Automated NLP systems [202] have been created to do this work effectively and accurately, much like how people digest small bits of text. Sentiment analysis has also received academic attention, which has sparked many studies looking at how it can be used to identify violent language [203]. Several NLP approaches, such as sentiment analysis [204], topic modelling [205], and identification of named entities [206], are frequently used to detect violence in text. These methods are essential for identifying occurrences of aggressive or violent language and classifying text as violent or non-violent.

By detecting words, phrases, or expressions related to aggression, anger, or harm, violent language can be detected by sentiment analysis, which examines the sentiment or emotional tone of a document. The discovery of underlying themes or topics in a text is made possible by topic modelling, which makes it possible to identify topics related to violence [207]. To gain insight into potentially violent circumstances, named entity recognition focuses on identifying particular entities such as persons, organisations, or locations referenced in the text [208]. These NLP approaches make it feasible to automatically analyse and classify text according to whether it is violent or not. This has important implications in several areas, including recognising hate speech [209], reducing cyberbullying [210], and helping law enforcement agencies identify threats [211].

It is crucial to remember that there are limitations [212] to detecting violence just through text analysis. Automated systems face difficulties when dealing with sarcasm, cultural nuances, and context. Therefore, improving the precision and efficacy of violence detection in text requires ongoing development of NLP approaches and incorporation of domain-specific knowledge. The proliferation of natural language texts calls for the employment of automated NLP systems to rapidly and effectively disseminate knowledge. These methods help to detect hostile or violent emotions and classify the material appropriately. To overcome the difficulties given by contextual comprehension and cultural variations and

ultimately improve the ability to detect violence in text with more precision, ongoing research and development in NLP are crucial.

The research surrounding sentiment analysis is focused on detecting negative and positive sentiments in data collected from social media platforms such as Facebook, Instagram, and Twitter, being the prominent sources [118, 213, 214]. Sentiment analysis has gained increasing recognition in the literature, as the analysis of an individual's emotional state and its dynamics can provide research with cues that could be used for predicting personality and speech patterns. These predictions can be used as a form of violence detection in different scenarios. Sentiment analysis has become more mature in the recent decade [120] and the most widely deployed classification techniques were SVM, Naive Bayes, and Maximum Entropy, which are based on the word bag model. However, the word bag model disregards the sequence of words in a sentence, which can have a significant effect on the meaning of the sentence, as well as change the sentiment, as discussed in the survey carried out by [121].

The detection of violent and offensive language detection is conventionally classified into specific types; such as the detection of bullying as seen in [122], identification of aggression [58] and hate speech identification [123]. NLP has been applied to a great extent in studies that involve sentiment analysis to exploit the syntactic lexical features of phrases and sentences to detect offensive language [124].

As NLP becomes increasingly popular for automatic detection of hate speech and abusive language, common patterns can be seen. Initially, data goes through pre-processing to gain useful insights and clean the text of irrelevant information using methods such as removing punctuation, stopwords, tokenisation, parts of speech tagging, and lemmatisation [125]. Machine learning-based classifiers were used [215, 216] to detect abusive language. However, existing word embedding methods [217], which use a limited window size, cannot exploit semantic information in the global context. In addition, such an algorithm transforms a word into a stable vector. As a result, the vector is unable to accurately represent its context at different locations.

4.3 Experimental Setup

The experimental setup is based on the methodology defined in Chapter 3, and allows the design and development of the model architecture, as well as testing the accuracy against basic models and other fusion attempts. For this part of the study, the experimental setup involved several Python libraries, including:

- Regex: used to work with regular expression and to remove symbols from text
- NumPy: used for numerical operations on arrays and matrices
- Pandas: used for data manipulation and analysis
- TensorFlow.keras: used for building deep learning models
- Matplotlib.pyplot and Seaborn: used for creating visualisations of the data
- Sklearn.preprocessing: used for pre-processing the data before training the models

For the initial step, the dataset went through the pre-proccessing procedure by removing all unwanted symbols, stop words, multiple spaces, single-character removal, punctuation, and numbers. The extraction of features was then required, which meant that the LIWC features and BERT features were chosen to be fused together to provide improved accuracy. The LIWC features were carefully selected through a principal component analysis (PCA) and those features were evaluated using statistical models, including the Random Forest Classifier and the K-Nearest Neighbours Classifier. LIWC features are based on linguistic and psychological theories, providing insights into emotional and cognitive content of the text, and the BERT features are based on NNs, capturing the context and meaning of words in the text. For this purpose, the BERT features were also extracted and then combined with the LIWC features for better reliability and accuracy.

4.3.1 Text Pre-Processing

Text pre-processing in NLP involves applying various techniques to text data in order to make it ready for use in NLP tasks. This may include splitting the text into individual words or punctuation marks (tokenisation), reducing words to their base form (stemming or lemmatisation), and removing common words that are not meaningful in the context of the task (stop word removal). Other pre-processing steps included include lowercasing the text, removing special characters, and ensuring that the text data are in a consistent format. Pre-processing is a crucial step in NLP as it helps to ensure that the text data is ready for use in downstream tasks.

4.4 Text Processing

Many text processing methods were considered as part of the project, and it was necessary to consider all options, including context-specific language models such as Universal Language Model Fine-tuning (ULMFiT) [218], OpenAI Generative Pre-trained Transformers (GPT) [219], Embeddings from Language Model (ELMo) [217] and BERT [220]. The work presented in this thesis relies on BERT, which provides a strong baseline, and further modifications to its standard network structure are performed to incorporate target information. The processing approach therefore followed the method of testing multiple different models and using the results to evaluate which method to continue with. Each NLP model and method was then used compared to existing and known methods to detect violent language.

4.4.1 Bidirectional Encoder Representations from Transformers

BERT is a language representation model developed by Google [112]. BERT was designed to allow bidirectional representations of unlabelled text to be performed, by considering the context of both right and left content across each layer. Taking into account the predicted content at each stage of the process. BERT is based

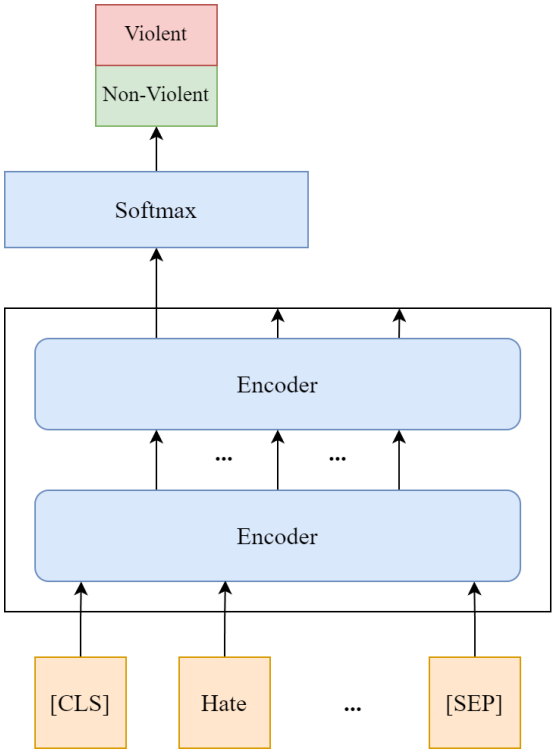

Figure 4.1: An illustration of the BERT process, provided in the context of binary violent language detection and classification. In this example, the input utterance is 'hate', which would be labelled as violent.

on a transformed architecture, which is a NN that was developed to work with sequential data, for example, text in the NLP scenario as presented in Figure 4.1. The transformer architecture enables the relationships to be captured between all tokens as opposed to just neighbour tokens. BERT has been found to be an effective method of working with NLP scenarios, for example, in the initial research conducted on BERT it was found to score highly in the range of tasks compared to state-of-the-art results, with a General Language Understanding Evaluation (GLUE) score of 80. 5% (7. 7% improvement), Multi-Genre Natural Language Inference (MultiNLI) precision of 86. 7% (4. 6% improvement), Stanford Question Answering Dataset (SQuAD) v1.1 question test F1 score of 93.2 (1.5 improvement), and SQuAD v2 test at 83.1 (5.1 improvement) [112].

BERT uses a tokenisation algorithm to split the text into a sequence of tokens, while there are various options for this, the WordPiece tokenisation method is

used within the context of previous research [112]. The tokenisation algorithm is initialised by representing the input text as a sequence of characters as presented in Figure 4.1, the algorithm then merges the most frequent relationships into new character sets until the iteration size is matched. The purpose of tokenisation is to computationally represent the text in a token format that can then be used in BERT. In the case of the research completed in this project, an existing dataset was used in this training process, in addition to a higher weighting of the words collected in Chapter 3. The WordPiece tokenisation algorithm can use maximum likelihood estimation (MLE) to find the maximum value/largest predicted probability of a token-token connection within the overall algorithm. The MLE formula can be represented as follows:

$$MLE = \arg\max_{S} \prod_{i=1}^{n} P(w_i|s) \tag{4.1}$$

where s is the token unit, n is the number of observations, i representing the value of the input sequence, P is the probability function and w is the word in the input sequence,

The BERT model is based on a transformer architecture, where self-attention is used to capture the relationship between each token in the sequence. After the completion of the tokenisation task, the self-attention equation enables BERT to focus on the most relevant input tokens. The BERT input sequence can be represented as a sequence of vector embeddings: $v_1, v_2, ..., v_n$, which are processed by the individual layer processes. When used as part of the self-attention layer, the embeddings can be used to ensure that BERT can focus on the relevant parts of the input. The formula for calculating the scaled dot-product which enables the resulting scores to be a weighted sum of the self-attention layer after applying a softmax function, can be represented as:

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QKT}{\sqrt{d_k}}\right)V, \tag{4.2}$$

where Q represents the query that is an input token, K is the key, V is the value, T is the transpose operator and D_k represents the dimensions of the keys and values from k and v [221].

The pre-training aspect of BERT is applied on a large corpus of text data split using two main tasks. The first task is masked language modeling (MLM) where masks are randomly applied to some tokens of the corpus and next sentence prediction (NSP), where the next token of a sentence is predicted based on the text corpus. This pre-training stage is an integral part of BERT, used for curating language representations that can be used for more contextual tasks. For example, in the context of this research project, this task is the detection of violent language from conversations.

MLM is used to randomly mask the input tokens from the initial stages of the model, MLM is then used to predict the original values of each token segment. During this training process, a [MASK] is placed in the token position during this training stage. The objective of this training stage is to increase the likelihood that the [MASK] tokens will be accurately predicted as part of the training process. The formula for maximum likelihood estimation used in the context of predicting a singular original value of a masked token based on the previous word can be presented as follows:

$$P\left(w_n|w_{n-1}\right) = \frac{C\left(w_{n-1}w_n\right)}{\sum_w C\left(w_{n-1}w\right)}, \tag{4.3}$$

where w is the input sequence, n is the number of the input sequence, $n-1$ is the previous word in the input sequence, and C is the computation of the bigram [222].

The final stage of this pre-training aspect is the NSP, during which BERT is tasked with two input sentences and attempts to predict if the two sentences are contiguous from the input text or not. The purpose of this task is to assist BERT in understanding the complex relationships between sentences during the pre-training stage of the model. By employing NSP, two sentences, denoted as A and B, are presented, with one of them being accurate and the other inaccurate. These classifications can be leveraged to establish sentence relationships for downstream tasks.

The final stage of BERT before the classification layer is the fine-tuning process. During the fine-tuning stage, the pre-training layers are fine-tuned for a task-specific purpose, which in the case of this research is violent language and

conversation detection. The fine-tuning process, therefore, largely depends on the purpose of the work and the dataset available. For example, some work may consider the use of sentence-pair classification, where a pair of sentences are matched in a similar method to the previously mentioned processes; this however requires a large-corpus of sentence:sentence classifications and does not match the data format used in this project. The fine-tuning in this project therefore focused on a NN approach, which was used to assist in classification of violent language during the project; the two methods used are presented in the remainder of this sub-section:

4.4.1.1 Long Short-Term Memory

LSTM is a type of NN which uses gates and cell states to solve the long term dependency problem and gradient issues in RNNs. LSTM networks therefore use cell states to provide memory, meaning the model can remember previous data points. LSTMs are made up of three main gates, the input gate, the forget gate, and the output gate. The input gate provides the input sequence for the other gates to then use, the forget gate removes unnecessary information from the cell state, and the output gate is used as an activation for the final output values of the model. For each gate, the cell states use the input and forget memory to then determine the output gate. The LSTM gates can be represented as follows:

$$\begin{aligned} i_t &= \sigma\left(w_i\left[h_{t-1}, x_t\right] + b_i\right), \\ f_t &= \sigma\left(w_f\left[h_{t-1}, x_t\right] + b_f\right), \\ o_t &= \sigma\left(w_o\left[h_{t-1}, x_t\right] + b_o\right), \end{aligned} \tag{4.4}$$

where i represents the input gate, f represents the forget gate, o represents the output gate, w represents the weight, h represents the previous output, x is the current input, and b is the bias for each gate [223].

4.4.1.2 Convolutional Neural Network

A 2D CNN was also an option to be used for the BERT fine-tuning, the 2D CNN was used to enable the processing of text-based data to reduce overall dimensionality of the output vectors. CNNs have traditionally been used for the

purposes of computer vision tasks; however, sometimes are used successfully for NLP scenarios such as classification or sentiment analysis. For NLP purposes, text segments are treated as a 2D image, where each row of pixels represents a sequence of tokens, and each column represents a dataset dimension. This 2D image format can then be used to extract features from text segments, enabling the representation of rich semantic details from the text.

The combination of a CNN and a BERT transformer model can provide some advantages to the overall algorithm which includes the improved feature extraction through capturing details of both local and long-range patterns in the dataset, the improved generalisation of the output data by capturing patterns across multiple segments, and improved interpretability of the resulting data through reducing the complexity of the BERT model.

4.4.2 Linguistic Inquiry and Word Count

To extract a rich amount of linguistic detail from the features of the dataset, the 2015 LIWC [224] psycholinguistic lexicon package was considered. The purpose of the LIWC package is to differentiate the semantic-syntatic patterns and the different contextual information from streams or collections of text. Based upon the context of this project, this would be focused on the themes relating to emotion and violence. A selection of the potential dataset classifications that could be used in the scenario of detecting violent language in conversations can include:

- Personal Pronouns: Includes the use of first person, second person and third person pronouns such as I, them, her, him, they, their, etc [224]. Personal pronouns could support the detection of multiple people being included in the conversation or conversations about another person occurring.

- Negations: Includes the use of denials and disproving language segments; some examples could include: no, never, not, and nothing [224]. Negations could be used within the proposed algorithm through the use of disproving statements and other language context.

- Positive Emotions: This includes the use of positive emotional language, including: love, nice, like, happy, sweet [224]. The positive emotions are

part of the affective processes sub-category within the LIWC psychological processes category.

- Negative Emotions: In a similar sense to negations, but in relation to emotive language, some examples of this could include: hurt, ugly, hate, nasty, and ugly [224]. Negative emotions are further classified into anxiety, anger, and sadness, all of which provide further classification context to the language input. The negative emotion category is also part of the affective processes category within the LIWC psychological processes category.

- Biological Processes: This category relates to biological process being identified in the text, across a broad range of categories including health, sexual, ingestion, and body. Some examples of these words could include: eat, blood, pain, clinic, flu, pill, spit, and hands [224]. Due to the nature of the words in this category, it could prove to be useful in a domestic violence or violent language detection algorithm.

- Perceptual Processes: The perceptual processes also contains multiple subcategories, which includes dictionaries of words relating to seeing, hearing, and feeling. Some example words from this category are look, heard, feeling, view, listen, touch, and view [224]. The words used in the category enable words related to perception to be detected from the input language text.

- Swear Words: The swear word sub-category is part of the informal language category. The swear word contains a dictionary of swear words that are identified as part of the LIWC package.

While the list above is only a non-exhaustive sample of the potential language categories that LIWC generates which could be relevant to violent language and conversation detection, it should not be taken as a complete understanding of the context of a conversation, during the initial phase of the project these language categories were used as assumptions which were then further analysed using PCA analysis and a bidirectional long-short term memory (Bi-LSTM) model. The models and known categories could then be weighted as necessary, and other categories were completely removed to reduce the overall dimensionality of the

overall model. The package also provides summary categories for the variables and dimensions that could be useful across a range of contextual use cases; these include summary language variables such as:

- Authentic: The authenticity algorithm is a summary measure relating to how authentic the input text is deemed to be by LIWC. The authentic measure is related to how self-monitoring an individual is, for example, in the context of being socially cautious [225].
- Emotional Tone: The emotional tone is a summary variable used by LIWC to provide a single variable for the positive and negative tone categories previously identified. A number below 50 suggests a negative emotional tone, and a higher number suggests a positive emotional tone [225].
- Analytical Thinking: Analytical thinking is another summary measure used by LIWC based on functional words; analytical thinking is a summary considering if more formal words are used. A low-scoring input text would suggest a more personal language classification, while a high ranking would suggest that a more formal language style should be used in the input text [225].

LIWC is provided as a package, which means that input and output data is generated using either the software or the website. For the purpose of this project, the software package was used, which enabled the import and classification of CSV files based on all the categories mentioned in the article "The Development and Psychometric Properties of LIWC2015" [224] article. LIWC can therefore be used to effectively analyse the contextual details of input text, including analysing summary methods such as personal language or emotion-based language. Although the contextual categories identified by LIWC do not provide a complete view, the methods used could effectively support the identification of violent language in domestic settings and conversations by adding contextual variables to the model.

4.4.2.1 Bidirectional Long-Short Term Memory

Bi-LSTM is a method of having an LSTM that has the sequence go in both forward and backward directions. In the bidirectional model, the data flows in both directions (forward and backward) to ensure that the information in both directions is stored and used. The Bi-LSTM provides an effective method of working with forward and backward information within NLP and speech recognition tasks. The Bi-LSTM was therefore used in addition to LIWC to ensure that the model was bidirectional and consider both forwards and backwards data formats.

4.4.2.2 Principal Component Analysis

One of the key steps in data mining and knowledge discovery is feature extraction, with the purpose of selecting the most suited features from a source with high dimensions to allow for more improved accuracy and reduced dimensionality at the classification stage of the process. To supplement the extraction of linguistic features using LIWC, a PCA was performed to distinguish the most contributing factors in the dataset. PCA is a type of dimensionality reduction that enhances the interoperability of large multidimensional data by summarising the content in large data tables, such as those generated during the LIWC processing. Figure 4.2 presents the full set of data generated by the LIWC processing, before a PCA took place. Figure 4.3 presents the highest contributing variables which include parameters such as functional words (to, very, it), analytical thinking, and common verbs (carry, think, go) as the most contributing dimensions post-PCA processing. Through the use of PCA, it was possible to reduce the dimensionality of the dataset, while preserving the variance and significance of the contributing variables. This reduction in the dimensionality of the data is useful given the purpose of the project to implement the models on edge devices.

4.4.3 Global Vectors for Word Representation

GloVe is an unsupervised learning algorithm for the production of high-quality word embedding. It is based on the occurrence matrix of words in the corpus, and the resulting embedded captures the semantic relationship between words.

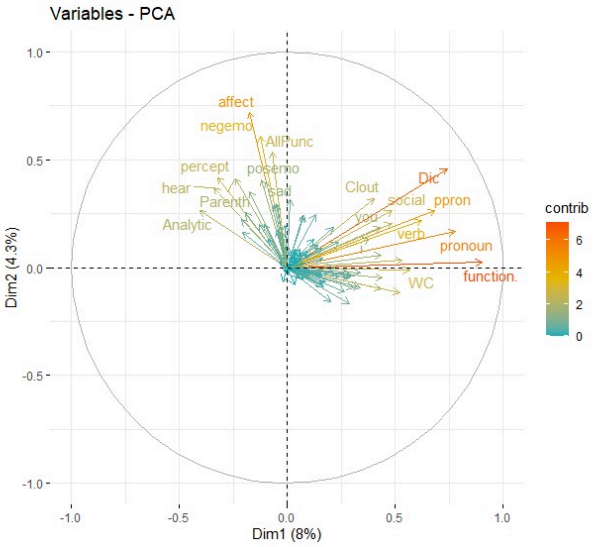

Figure 4.2: The values of the LIWC features prior to the PCA being conducted. The values and labels present all of the contributing variables in the LIWC feature collection.

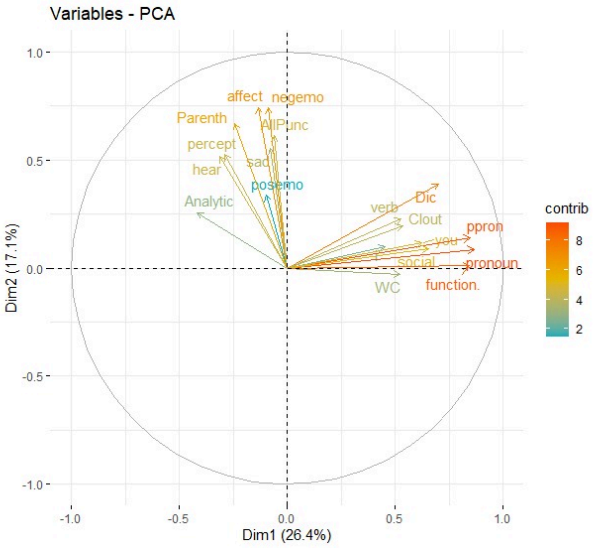

Figure 4.3: The values of the LIWC features after the PCA has been conducted on the dataset. The remaining variables and labels are those that contributed the most to the features.

The GloVe algorithm is constructed of a matrix of word-to-word co-occurrences, where each element of the matrix represents the number of times a word ap-

pears in the context of another word. Then the model factors the matrix of co-occurrence using a weightless method of least squares to obtain low-dimensional vector representations for each word. The objective function of the GloVe model is as follows:

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f\left(X_{ij}\right) \left(\mathbf{w}_i^T \mathbf{w}j + b_i + b_j - \log Xij\right)^2 \tag{4.5}$$

In this formula, V is the vocabulary size, X_{ij} is the number of times word i appears in the context of word j, $\mathbf{w}_i$ and $\mathbf{w}j$ are the word vectors for words i and j respectively, b_i and b_j are the bias terms for words i and j, and $f(Xij)$ is a weighting function that reduces the impact of very frequent co-occurrences.

4.5 Model

For the models selected as part of the NLP section, the BERT and LIWC methods were developed separately, before being fused as part of Chapter 5. The models were both developed based upon the initial understanding of the model presented in the previous section, before being fine-tuned and analysed for the purposes of presenting results as part of this Chapter.

4.5.1 Bidirectional Encoder Representations from Transformers

The BERT model implemented as part of this work is developed following the initial pre-trained model used in BERT instances within Keras. The BERT pre-trained model uses the `bert-base-uncased` model to support this process, this model is already set up using three layers including a main BERT layer, a dropout layer, and a classification layer. This model imports over 109 million existing BERT parameters which can then be customised based on the specific contextual needs of the scenario, in the case of the work presented in this thesis, this is violent language detection from conversations. The custom dataset used in this model is the dataset generated as part of Chapter 3, and then pre-processed using the

methods identified previously in this chapter. The imported BERT model before the additional dataset word weightings has a layer organisation as follows:

```
Layer (type)                 Output Shape              Param #
bert (TFBertMainLayer)       multiple                  109482240
dropout_37 (Dropout)         multiple                  0
classifier (Dense)           multiple                  1538

Total params: 109,483,778
```

Using this dataset, custom word weightings can be generated using the more curated set of data relevant to violent language. For this purpose, the dataset is first extracted and then formatted to match the imported data source, specifically the main BERT layer. The layer is then tokenised to form a combination of the required tokenisations from both the imported dataset and the pre-trained model. The tokenisations being completed enables the model to use the tokens as text representations. Following this, a 70/30 train/test split is used on the combination of both datasets, with processing focused on the imported data.

At this stage, the pre-training of the model is performed, where the normal BERT processes of MLM and NSP occur with the dataset provided. This provided dataset includes both the pre-trained model and the additional parameters imported from the custom dataset. The model uses the input IDs to link the language tokens, before applying an attention mask. Additional layers can be added at this part of the process; in this case, a LSTM and CNN are applied to the model in different versions, meaning that during the results and analysis stage of the project, two individual scores will be presented. The LSTM and CNN layers are added before the processing is completed on the model, and are added as layers to the above layout. Alternative methods were not included due to existing concerns regarding the large dimensions of the dataset; for example, next sentence prediction (NSP) was not included due to the complexity of this fine-tuning method; instead, more lightweight layers were chosen.

In the BERT model training stage, the optimal model was explored for the dataset, research explored the following hyperparameter spaces: activation functions of rectified linear units (ReLU) or linear, Adam optimiser with learning rate

(6.25e-3, 6.25e-4, 6.25e-5, 6.25e-6), and number of epochs (1, 5, 10). To reduce over-fitting, a regularisation lambda of λ=0.01 is utilised and batch normalisation is applied after each layer. The selected model parameter settings were based on the initial results collected, each model is trained using the Adam optimiser (with a learning rate of 6.25e-4), and a batch size of 32 for 1 epoch. The activation function for both datasets in the model was selected to be ReLU.

Once the model has been trained, the resulting BERT vectors can be used for the purposes of extracting further information, or in the multimodal fusion process as presented in Chapter 5. The output of the BERT model is a vector matrix with a shape of (30522, 768), which presents the word weightings based on the pre-trained and then completed model. The vectors enable these generated weightings to be further applied in other work. The output extract presented below is a sample of these vector word weightings:

```
[[-0.01018257 -0.06154883 -0.02649689 ... -0.01985357 -0.03720997
  -0.00975152]
 ...
 [ 0.00145601 -0.08208051 -0.01597912 ... -0.00811687 -0.04746607
   0.07527421]]
```

To capture the metrics of the model, Keras is used. Keras enables the metrics of the model to be captured effectively while also allowing the reporting of the accuracy, loss, and F1 score of the model. This is necessary to allow the results of the model to be effectively reported within this thesis.

4.5.2 Linguistic Inquiry and Word Count

The model uses LIWC embeddings which are generated for each word, followed by the PCA performed earlier in this chapter. LIWC embeddings are generated for each word through importing the dataset into the LIWC application, which enables the individual values to be generated for each word embedding. The LIWC application exports the embeddings for each word as a CSV file, enabling the file to be imported into other applications and other use-cases. The LIWC dataset is imported into the Python runtime environment to enable the features

to be used in the model. The embeddings at this stage are still static values; however, once imported, the values and rankings can be used by the custom code to support the development of custom models. LIWC enables relationships between words to be recognised for each point available within the dataset. Due to the PCA being performed fewer overall LIWC values are provided because of the high dataset dimensionality.

When combined with other fusion modalities, LIWC can be used to support the resulting values. This is performed by importing the dataset and then splitting the dataset using the same ratio as existing BERT embeddings to ensure that the word and value pairs match. Using the matched word embeddings, it is then possible to further augment the results of the processing by including the LIWC categories to enhance the features from existing sources. When processed, the LIWC model is combined with a Bi-LSTM to enhance the capabilities of the model when combined with existing data sources.

The standalone LIWC model produces no overall result; therefore, for the process of this chapter, LIWC is combined with other methods to generate usable and comparable results of the algorithms applied during the work. LIWC is combined with K-Nearest Neighbours and Random Forest Classifier. Although it is not expected to produce meaningful results with these methods, these can be used as baselines to compare future use of the model with. During this stage, the model is still using the previously identified Bi-LSTM, before then being applied to each of the baseline measures.

Implementing K-Nearest Neighbours in the model first involved selecting the points, based on the comparative distances between the values based on the similarity between two-word embeddings using the Euclidean distance. Using the distance calculations of the word embeddings, it is possible to find the nearest neighbours of each value, based on the distance metric. The neighbours can be selected on the basis of the K training points and the shortest distances. This can then be used for classification, where the K-point is assigned to the expected classification task. Similarly, implementing the random forest classifier uses a random subset of the training word embeddings and determines the number of trees based on cross-validation. To train the decision trees, the aggregation process is used before each decision tree is split according to the information gain

splitting criteria. The majority vote is then used by the model to classify the results of each tree.

The processing of LIWC is completed using the same values as the BERT classification process, also applying Adam for optimisation, and ReLU as the activation function. In addition, a batch size of 32 is selected over 1 epoch. The results of the LIWC and the baseline combinations are reported as F1 scores, which are captured using Keras.

4.6 Results

The results of this section are divided into three tables, Table 4.1 presents the values of the F1 scores of the standalone methods, Table 4.2 presents the LIWC baseline measures and Table 4.3 presents the F1 scores of the BERT model with fine-tuning.

Table 4.1: Results displaying the values of the features in the frequency domain when compared using different methods. The methods are compared using the F1 score.

Model	**F1 Score**
CNN + Glove	0.6370656
NN + Glove	0.6602316
LSTM + Glove	0.5945945

The results of the text processing were reported using the F1 scores of the models, using metrics that were captured in Keras. Due to the nature of the text processing and the ease of use of combining models and methods, various attempts were made to measure the highest accuracy models. Initially, combinations of traditional classification methods and individual methods produced relatively low scores, as presented in Table 4.1. The highest result for the Glove-based approach was the combination of NN and Glove, which achieved an F1 score of 0.66. This was in contrast to the two other methods attempted, which include the CNN and Glove score of 0.64 and the LSTM and Glove score of 0.59. The results were all within a range of .1 indicating that Glove had a varied impact on the model for

each implementation developed. Further fine-tuning of the Glove approach could be undertaken to improve these results, however based on the higher scoring F1 measurement this should focus on the NN combination approach.

Table 4.2: Results displaying the values of the LIWC method when combined with other baseline measures. The methods are compared using the F1 score.

Classification	**F1 Score**
LIWC + K-Nearest Neighbors	0.596401
LIWC + Random Forest Classifier	0.591259

Table 4.2 presents the resulting values of the LIWC classification when combined with baseline models to ensure a classification can be performed. Individually, the LIWC classification models did not perform well, with the LIWC and K-Nearest-Neighbour combination achieving a resulting F1 score of 0.60 on the classification task for the dataset. The results of the LIWC and the random forest classifier were also poor, achieving a lower resulting F1 score of 0.59. The nature of LIWC as a language analysis tool meant that the results reported using this approach were expected to be low, however the intention of the LIWC model is for this to be combined with other models for the purpose of integrating additional context and information. While BERT and Glove are both vector-based approaches, LIWC provides a contrasting word analysis tool, which could assist in enhancing the generalisation of the model.

Table 4.3: Results presenting the final F1 scores of the BERT and LSTM/CNN combinations. The methods are compared using the F1 score.

Model	**F1 Score**
BERT + LSTM	0.6602
BERT + CNN	0.6681

Finally, the results of the BERT model are presented in Table 4.3, the resulting values are presented in addition to fine-tuning the model using LSTM and CNN respectively, enabling an overall comparison to be made. Combining BERT with an LSTM produces a resulting F1 score of 0.66, highlighting the value as higher

than any of the previously reported results. The resulting value of the BERT and CNN combination is 0.67, which is the highest resulting value at any stage of the text model aspect of this chapter. The BERT and Glove models are similar, with the highest measurement for each of the approaches having less than a 0.01 difference in value.

4.7 Discussion

Table 4.4: A complete overview of the results collected throughout this chapter, presented as a comparison of F1 scores to enable a comparison of results to be formed.

Model	**F1 Score**
CNN + Glove	0.6371
NN + Glove	0.6602
LSTM + Glove	0.5946
LIWC + K-Nearest Neighbors	0.5964
LIWC + Random Forest Classifier	0.5913
BERT + LSTM	0.6602
BERT + CNN	0.6681

This chapter introduced the text processing methods used within the project to detect violent language from the previously curated violent language multimodal dataset. The best method of identifying violent language and conversation from prerecorded and pretranscribed audio was investigated for the purposes of implementing these features into a multimodal fusion model. For these purposes numerous considerations were needed, most notably the consideration of the features extracted and how the techniques would support other features was considered; for example, using a method that does not rely on word embeddings or a nonmatching set of word embeddings would not be effective when combined with a model such as BERT which does rely on such embeddings. This chapter presented a wide array of techniques for classifying violent language, such as

baseline models to be combined with LIWC features, methods of fine-tuning the BERT model, and weighted word embeddings and vectors alongside existing data sources.

A complete breakdown of the results collected within this chapter is presented in Table 4.4. Using the F1 score in NLP ensures that the results accurately capture true linguistic patterns without confusing them with irrelevant data. It helps avoid overlooking significant language features while preventing the misclassification of benign text, maintaining a balanced approach in detecting nuanced language use. The highest result was the combination of BERT and CNN, which achieved an F1 score of 0.67 in classifying violent language using the curated dataset. Compared to existing literature such as the context-free text model, the weighted text model and the sequence text model at F1 values of 0.5, 0.44, and 0.67 respectively [226]. The BERT and CNN model does not achieve the same score as the specific modelling task performed in previous work, which reported an F1 score of 0.77 [227], possibly due to the task-specific modelling performed in the work. BERT values, when combined with CNN or LSTM, perform well compared to previous results reported in the table, with fine-tuning improvements that could support further improvements to the F1 score.

The initial baseline methods performed well, specifically the methods that included a CNN, with the CNN and Glove combination reporting an F1 score of 0.64 as presented in Table 4.4. A NN was better than CNN when combined with Glove with a score of 0.66, while the LSTM based approach performed poorly compared to an F1 score of 0.59. The CNN model and the NN + Glove model performed well compared to the highest values of BERT and CNN, with the values highlighting the effectiveness of Glove when considered as part of the work. These initial measures also highlight the effectiveness of standalone methods without the need for additional features or processing to be applied. These models could be returned at a future time to investigate how such methods could further improve the BERT and CNN score and support improved feature extraction.

The results of the LIWC baselines were expected to be poor, and reported poor F1 values for the classification tasks; the LIWC values were not the end result of the LIWC processing, with the PCA values reported previously in the chapter forming part of the overall model, in addition to the LIWC processing

being more effective when considering the multimodal nature of the next stage of the work. Despite this, the baseline LIWC values reported in Table 4.4 would not be effective for use in an overall system, due to the high margin of error afforded by the values reported. Improvements to the dimensions of the dataset would probably improve this overall value; however, the nature of LIWC is not intended to work well at a stage with a single modality included.

In comparison to recent literature using BERT and LIWC, the results of the model developed during this chapter are positive. Similar work has previously used BERT and LIWC for the purposes of detecting propaganda from news articles, attempting to implement BERT and LIWC for new approaches using BERT features, part-of-speech features, and LIWC features with a logistic regression model [228]. The work [228] explored the integration of a propaganda spans' detection framework with the BERT model achieving an F Measure of 38.88, the article also presents the values of 38.51 for the BERT, part-of-speech, and LIWC results. The approach proposed by the authors [228] contrasts to the approach used within this thesis in both dataset and purpose, although the results of this work are encouraging through the F1 score of 0.67. LIWC and BERT for text has also been used previously, although the authors exploring BERT and LIWC did not fuse the results of the two models [42]. The method proposed in an article on predicting communication behaviours during couple conflicts [42] also integrates an SVM approach with BERT and LIWC, however, the approach used does not integrate both models combined, but concludes a similar outcome of the BERT-only model performing better than the LIWC-only model. The architecture reported in this thesis contrasts with similar approaches [42, 228] through the complexity and features used in the model, highlighting the potential future capability of the highly complex models introduced in this chapter.

The BERT model provided novel research scenarios that needed to be investigated over the course of the project, for example, the initial parameters of BERT included 109,482,240 values, while the custom-curated dataset only had 1538 parameters. The overall model needed to be weighted effectively, to avoid task-specific modelling, these values were carefully weighted to ensure that new transcription segments were being weighted accurately. The difference between the two combinations is minimal, although still a relatively low score compared

to previous work, such as work investigating the detection of depression using an audio-text approach with a LSTM NN which reported an F1 score of 0.77 [226]. The model showed improvement in some reported values, most notably text-only models that generally performed poorly when working with multimodal data sources. Interestingly, the results for the text-based modality are much lower than those reported later in the audio fusion approach; this may be due to the limited number of transcriptions available in the current dataset, which is hoped to be expanded in future studies.

Overall, the methods presented in this chapter provided some novel discoveries in terms of known methods of text processing in addition to combinations of techniques such as BERT and LSTM, BERT and CNN, and NN and Glove. The models and results presented in this chapter provide sufficient evidence of the problem of violent language detection from non-social media sources, especially considering the transcription problems during this process. The architecture for the BERT and LIWC combination conducted at present does not yet improve the accuracy of text-based language models in similar contexts, however, the potential of the work highlights the variation of individual BERT and LIWC models, Therefore it is necessary for the next stage of research to focus on the implementation of a multimodal fusion approach, combining the reported values of both text and audio models into a single F1 score.

4.8 Conclusion

This chapter presented text-processing models using NLP to detect violent language from conversation transcriptions. This chapter has considered the implementation of a variety of models, including Glove, CNN, LSTM, Bi-LSTM, BERT, and LIWC to determine which method works best for analysing potentially violent language. The models that performed the best outside of the baseline CNN were those that used BERT, with these models achieving an F1 score of 0.67 and 0.66 for CNN and LSTM results, respectively. Although LIWC was considered for the categories it provides, with baseline models performing poorly, but expected to improve when combined with other features. The models presented in this chapter were not overly effective; however, it is expected

that the models presented in this chapter will become more effective upon fusion with audio-based features. Future work should focus on methods for fine-tuning BERT, which could potentially be further improved by applying other methods during the fine-tuning stage.

Chapter 5

Audio Inference and Multimodal Fusion

5.1 Chapter Overview

In this chapter, the use of audio inference and diarisation techniques for crime prevention will be explored. The chapter begins by defining audio inference and diarisation and discussing their importance in the context of crime prevention. Following that, the current state-of-the-art in audio inference and diarisation is examined, covering the numerous techniques and algorithms that have been developed, as well as their significant strengths and limitations. The field's challenges and outstanding research problems are also mentioned.

The particular uses of audio inference and diarisation for crime prevention are then examined, including the use of these approaches for identifying suspects, detecting and analysing conversations, and monitoring and analysing individual movements. There is also a discussion of the ethical and legal aspects that must be taken into account when applying these tactics to prevent crime.

Potential future developments in audio inference and diarisation and how they may be used to enhance crime prevention efforts are considered. This includes a discussion of the potential for integrating these techniques with other technologies, such as video surveillance and facial recognition, to develop more comprehensive and effective crime prevention systems.

5.2 Background

The field of emotion recognition using audio-based technologies has gained increasing attention in recent years, leading to an increasing number of research studies in the field [229–233]. For example, the work in [184] proposes a Bi-LSTM (referenced as BLSTM) with attention model to classify four different emotions including "anger", "excitement", "neutral", and "sadness" from the IEMOCAP dataset, which includes a range of dyadic sessions where actors perform improvisations or scripted scenarios specifically selected to elicit emotional expressions. To do so, a silence removal signal processing step followed by the extraction of a 34-dimensional feature vector composed of a range of time domain, frequency domain, MFCCs and Chroma-based features is implemented. The results reported outline a maximum classification accuracy of 70.34%. Using the same dataset, [183] proposes a self-attention CNN-BLSTM classification model fed Mel spectrograms of the corresponding audio files to classify between the four emotions mentioned above. The above methodology achieves an 81.6% classification accuracy. In [234] a multi-channel CNN fed with Mel-spectrograms and phoneme embeddings generated by the word2vec model introduced in [235].

The multimodal fusion approach has been used for emotion recognition using audio, video, and physiological data [236]. Sahu proposes a feature engineering-based approach to address speech emotion recognition, similar to the attempted approach. They first trained all their audio and text features separately and then fused them to the concatenated feature vectors [237]. Chen et al. approached multimodal sentiment analysis by incorporating audio and text. Their strategy was to use both multi-feature fusion and multi-modality fusion to improve the accuracy of audio-text sentiment analysis. Their focus was solely on general sentiment rather than the detection of hateful or negative sentiment, and therefore used the CMU-MOSI dataset [124]. Pereira et al. [238] utilises audio, textual, and visual cues extracted from news videos, which applied state-of-the-art computational methods to automatically recognise emotion from facial expressions; The authors use a dataset that contained 520 annotated news videos from three famous TV newscasts and reported an accuracy of up to 84% for the sentiment classification task [238].

5.3 Fusion Modalities

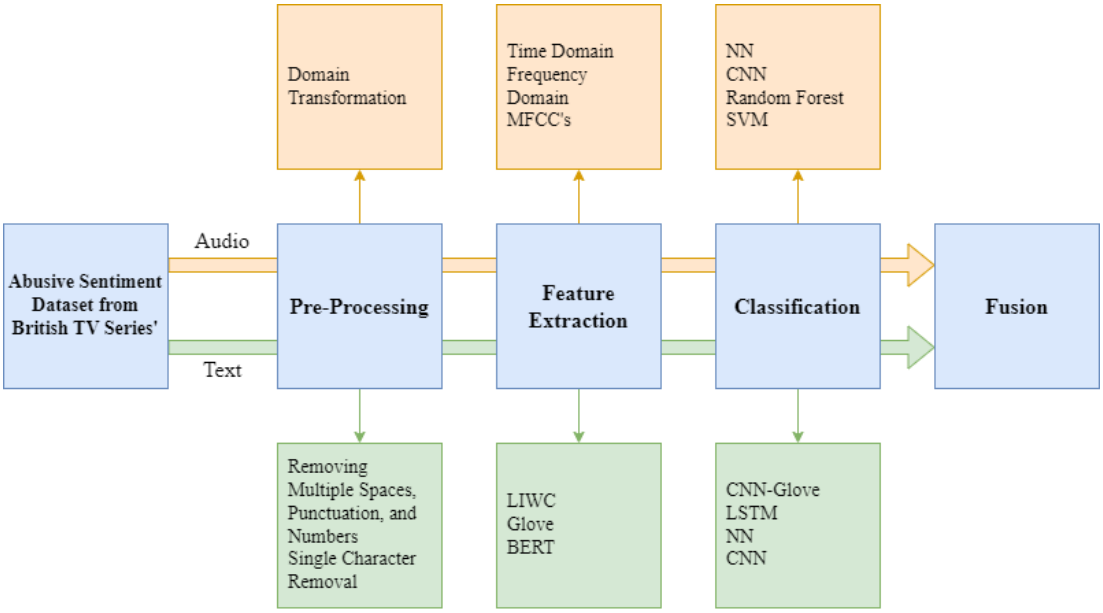

Figure 5.1: The overall architecture of the multimodal fusion approach, and the individual methods used across each modality. The diagram first introduces the dataset produced, followed by the pre-processing, feature extraction, classification, and fusion for both audio and text modalities.

Before determining the use of which modalities, it was understood that the NLP approach was effective, but needed further enhancement. As identified in Chapter 2 and Chapter 4, the potential improvement could come from multimodal data, applying the work beyond a single text-based modality to achieve more effective results. Although NLP is good at analysing and interpreting text data, it is not always sufficient to provide comprehensive knowledge of a situation. In this case, the combination of other data streams, such as audio, allows additional context, which can aid in improving the accuracy of the analysis as presented in Figure 5.1.

Fusion techniques have several benefits and, in this case, ensure that emotional signals such as pitch, tone, and volume cannot be missed through a multimodal approach. These can easily be overlooked in mere audio transcriptions if the abuse is not explicitly conveyed through verbal profanity. For example, if someone

playfully swore for comedic effect, the NLP analysis would pick that up as abusive text and classify it wrongly; however, with the audio analysis, the tone and pitch would be picked up to avoid such an error. To provide another example, someone may say that they are 'fine' in a monotone voice, which would be interpreted as neutral or as a positive statement even though they are actually upset/distressed, which is clear by their tone. Likewise, an NLP analysis would also not be able to pick up any use of non-literal language such as sarcasm and irony that is conveyed through the tone rather than words.

There are other fusion modalities that could have been implemented, such as video, as other investigations [239] and [240] found that by using a combination of data streams, they were able to achieve higher accuracy. However, they used audio and video for the purpose of their studies which did not seem appropriate for a domestic setting. Therefore, by analysing the actual audio data instead of just the transcription it is possible to improve the overall accuracy of the classification as it may be possible to pick up on emotional cues and other forms of non-verbal communication that are not captured for the text analysis. This provides a more comprehensive understanding of the situation and conversation, eliminating any forms of doubt or misunderstanding of text data, and helps identify abuse or other interpersonal conflicts.

5.3.1 Natural Language Processing

NLP has been well known for its effectiveness in recognising patterns and relationships in large volumes of text data and has been shown to be valuable in a number of applications, such as online cyber bullying. Additionally, NLP is helpful in uncovering linguistic patterns that are symptoms of violence or abuse. Several studies have investigated the use of NLP approaches to identify examples of domestic abuse; for example, a study published in the Journal of Interpersonal Violence [241] analysed a large corpus of discussions on domestic abuse-related Internet forums using NLP. The authors [241] discovered that their NLP approach accurately identified posts containing indications of abuse, the study also acknowledged that there are limitations in solely relying on data that are based on text and suggested using other modalities.

NLP is a valuable tool to identify and analyse data from a large corpus of text and can enable sentiment analysis, topic modelling, and named entity recognition. The study [242] used NLP techniques to extract features from a dataset of 3 million ISIS-related tweets, which also contained metadata such as follower count and location. NLP has therefore been applied in a range of academic contexts, and could potentially be improved through a multimodal approach to support text classification. The BERT and LIWC models presented in the previous chapter will be used, due to the richness of the features extracted.

5.3.2 Acoustics

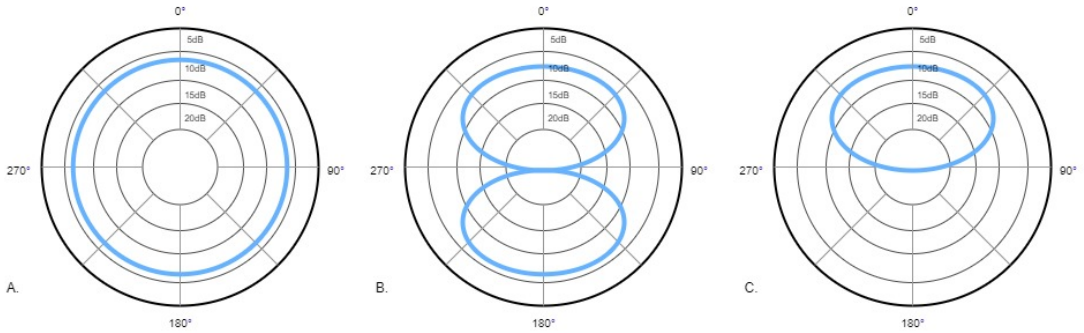

Figure 5.2: A demonstration of the three main types of microphone input patterns, with omnidirectional (left) for even sound quality, bidirectional (middle) for two directional sound quality, and unidirectional (right) for singular direction capture. In the context of a smart home microphone, an omnidirection would be able to capture sounds from across the room, while a unidirectional microphone would only capture the direction it is facing.

Audio processing combines the time domain, frequency domain, and MFCC into a single model to detect violent conversations on edge devices. To enable useful detection of audio patterns, the microphone of both the initial dataset and the final classification model should be considered as presented in Figure 5.2. For the purpose of this work, an omnidirectional microphone was selected. The time domain enables the examination of signal characteristics in the time dimension, which includes signal energy, signal duration, and amplitude. The frequency domain provides insight into the spectral properties of audio signals, which include

frequency components and the related distribution. By including by time and frequency domains, the model can process important and contextual information about the temporal and spectral characteristics of violent conversations, with the aim of improving the accuracy of detection.

MFCC is then included to transform audio signals into a set of cepstral coefficients that can process the unique and contextual feature of speech. The MFCC approach is widely used and has been shown to be effective in identifying languages and speech patterns; therefore, it is hoped that this processing method will improve the detection of high-level features of violent conversations and the language that occurs in them. This enabled the detection of the tone and emotion of the recorded segments, which improves the accuracy of the fusion model. Using a combination of these techniques enables the algorithm to take advantage of the strengths of each modality to capture a comprehensive, contextual, and nuanced understanding of what violent language occurs and violent conversations to support crime prevention efforts.

5.4 Audio Inference

The literature review and the resulting NLP results highlighted a turning point in the focus of the research project. This section explores the audio inference that was implemented in the development of multimodal fusion, highlighting the process, techniques, and algorithm design used to enhance the possibility of detecting violent conversations through multimodal fusion. By conducting a thorough analysis of existing methods and studies, it was determined that audio inference was possible to implement on edge devices. Furthermore, the inclusion of audio inference in the overall model had the potential to improve the knowledge of multimodal data fusion and the effectiveness of detecting violent language on the edge. Although the audio-based approaches presented are not state-of-the-art, the techniques used are effective for edge and mobile devices due to processing speeds. The remainder of this section explores audio inference.

5.4.1 Experimental Setup

The experimental setup follows that of the methodology defined in Chapter 3. The experimental setup enabled the design and development of the model architecture, as well as testing the accuracy against baseline models and other fusion attempts. The experiments were carried out on audio segments of the dataset, during which the linked audio files were used as the main source of data, along with the CSV violence ratings. This dataset had already been pre-processed as identified in Chapter 3 and therefore any noise and outliers had already been removed. The algorithm was implemented on Google Colab using Python, with the necessary parameters introduced within this section. The models were developed using the TensorFlow Keras library due to development speed and the opportunity to convert models to TensorFlow Lite for edge computation. Pandas, NumPy, and sklearn were also used as libraries for the processing and mathematical formulae used to process audio data. The resulting F1 scores were compared alongside existing literature to ensure that the algorithms are running effectively. Overall, this experimental setup ensures that the results are robust and reliable and allows for a fair comparison of the approach with existing methods.

5.4.2 Audio Processing

Speech recognition, or speech-to-text, is the computational process of identifying spoken words within an audio signal and preserving those words in readable text. To enable the accurate recognition of words and the subsequent translation to text, the audio files need to go through key pre-processing steps. These include the removal of unwanted background noise that does not correspond to human speech, as well as the segmentation of the audio signal and the grouping of homogeneous regions of the audio signal regarding the speaker identity (speaker diarisation). With this, the aim is to obtain N "noise-free" audio signals corresponding to the speech of the N speakers who are involved in a conversation.

The original sampling rate of the different audio signals that make up the proposed dataset is 22050Hz. However, lower sampling rates can lead to higher classifications, as the far end of the spectrum is not expected to contain information for speech-related applications. Given this, the performance of the system

was studied across different sampling rates (22,050Hz, 16,000Hz, and 11,000Hz) through the downsampling of the audio signals. Unlike other sensory signals or data formats such as inertial signals from inertial measurement units (IMUs), physiological signals, or images. The extraction of patterns from raw audio data typically involves the extraction of hand-crafted features to achieve competent performances. This is due to raw audio data already being in the form of time series data, whereas substantial information about the audio signals can only be revealed within the frequency domain. The proposed approach is a feature vector that incorporates a wide array of self-engineered features, revealing relevant information about the signal in the time and frequency domains, in addition to the MFCCs extracted.

5.4.2.1 Time Domain

The array of features calculated in the time domain can be defined as follows:

- Amplitude envelope:

$$AE_t = \max_{k=tK}^{(t+1)K-1} s(k), \tag{5.1}$$

  where t refers to the t^{th} frame, K is the frame size and $s(k)$ is the amplitude of the k^{th} sample of the signal.

- Amplitude envelope discrete derivative:

$$\Delta AE_t = AE_t - AE_{t-1}, \tag{5.2}$$

  where AE_t is the amplitude envelope of frame t and AE_{t-1} is the amplitude envelope of frame $t-1$.

- Root mean square energy:

$$RMS_t = \sqrt{\frac{1}{K} \sum_{k=tK}^{(t+1)K-1} s(k)^2}, \tag{5.3}$$

where t refers to the t^{th} frame, K is the frame size and $s(k)$ is the amplitude of the k^{th} sample of the signal.

- Root mean square discrete derivative:

$$\Delta RMS_t = RMS_t - RMS_{t-1}, \tag{5.4}$$

where RMS_t is the root mean square energy of frame t and RMS_{t-1} is the root mean square energy of frame $t-1$.

- Zero crossing rate:

$$ZCR_t = \frac{1}{2} \sum_{k=tK}^{(t+1)K-1} \left| sgn\left(s\left(k\right)\right) - sgn\left(s\left(k+1\right)\right) \right|, \tag{5.5}$$

where t refers to the t^{th} frame, K is the frame size and $sgn(s(k))$ is the sign of the amplitude of the signal at sample k.

- Zero crossing rate discrete derivative:

$$\Delta ZCR_t = ZCR_t - ZCR_{t-1}, \tag{5.6}$$

where ZCR_t is the zero crossing rate of frame t and ZCR_{t-1} is the zero crossing rate of frame $t-1$.

Following the computation of the above time domain features, basic descriptive statistics, including the mean, the maximum, the minimum, the standard deviation and the root mean squared from each of the features are calculated, leading to a 30-dimensional feature vector (6 features x 5 descriptive statistics) presented in Table 5.1.

5.4.2.2 Frequency Domain

Alongside the features extracted in the time domain, a number of features are extracted in the frequency domain. To do so, the time series x, corresponding to the audio signal, is converted into the frequency domain using the Short Time

Table 5.1: The 30-dimensional feature vector of the time domain features, represented as a table (6 features x 5 descriptive statistics). The details of the vector are presented as descriptive values of the feature set.

Feature	Mean	Max	Min	STD	RMS
AE	mean_ae	max_ae	min_ae	std_ae	rms_ae
delta AE	mean_delta_ae	max_delta_ae	min_delta_ae	std_delta_ae	rms_delta_ae
RMS	mean_rms	max_rms	min_rms	std_rms	rms_rms
delta RMS	mean_delta_rms	max_delta_rms	min_delta_rms	std_delta_rms	rms_delta_rms
ZCR	mean_zcr	max_zcr	min_zcr	std_zcr	rms_zcr
delta ZCR	mean_delta_zcr	max_delta_zcr	min_delta_zcr	std_delta_zcr	rms_delta_zcr

Fourier Transform (STFT) along with a Hann window as follows:

$$S\left(m,k\right)=\sum_{n=0}^{N-1}x\left(n+mH\right)w\left(n\right)e^{-i2\pi n\frac{k}{N}}, \tag{5.7}$$

where m refers to the m^{th} frame or temporal bin, k refers to the k^{th} frequency within the frequency bins, N is the frame size, H is the hop size and $w(n)$ is the Hann Window function applied to the n^{th} sample within a frame m. The Hann Window is given by:

$$w(k)=\frac{1}{2}\left(1-cos\left(\frac{2\pi k}{K-1}\right)\right),\quad k=1,\ ...,\ K. \tag{5.8}$$

Once the signal is converted into the frequency domain and the windowing function is applied to each signal frame, the following features are extracted:

- Band energy ratio (BER):

$$\Delta BER_t=\frac{\sum_{n=1}^{F-1}m_t(n)^2}{\sum_{n=F}^{N}m_t(n)^2}, \tag{5.9}$$

  where $m_t(n)^2$ is the power of the signal at frame t and frequency bin n, F is the split frequency and N is the highest frequency bin within a frame t.

- Band energy ratio discrete derivative:

$$\Delta BER_t = BER_t - BER_{t-1}, \tag{5.10}$$

where BER_t is the band energy ratio of frame t and BER_{t-1} is the band energy ratio of frame $t-1$.

- Spectral centroid (SC):

$$SC_t = \frac{\sum_{n=1}^{N} m_t(n)n}{\sum_{n=1}^{N} m_t(n)}, \tag{5.11}$$

where $m_t(n)$ is the magnitude of the signal at the n^{th} frequency bin and N is the highest frequency bin within a frame t.

- Spectral centroid discrete derivative:

$$\Delta SC_t = SC_t - SC_{t-1}, \tag{5.12}$$

where SC_t is the spectral centroid of frame t and $SCt-1$ is the spectral centroid of frame $t-1$.

- Spectral bandwidth (SBW):

$$SBW_t = \frac{\sum_{n=1}^{N} |n - SC_t|\, m_t(n)}{\sum_{n=1}^{N} m_t(n)}, \tag{5.13}$$

where $m_t(n)$ is the magnitude of the signal at the n^{th} frequency bin in frame t, N is the highest frequency bin within a frame t and SC_t is the spectral centroid at frame t.

- Spectral bandwidth discrete derivative:

$$\Delta SBW_t = SBW_t - SBW_{t-1}, \tag{5.14}$$

where SBW_t is the spectral bandwidth of frame t and SBW_{t-1} is the spectral bandwidth of frame $t-1$.

- Spectral roll-off:

$$\sum_{n=0}^{R_{n-1}} |m_t(n)| = 0.85 \sum_{n=0}^{N-1} |m_t(n)|\,, \tag{5.15}$$

where $m_t(n)$ is the magnitude of the signal at the n^{th} frequency bin in frame t, N is the highest frequency bin within a frame t and R_{n-1} is the roll-off frequency bin.

- Spectral roll-off discrete derivative:

$$\sum_{n=0}^{R_{n-1}} |m_t(n)| - \sum_{n=0}^{R_{n-1}} |m_{t-1}(n)|\,, \tag{5.16}$$

- Spectral Flux (SF)

$$SF_t = \sum_{n=0}^{N-1} s(k,i)s(k-1,i), \tag{5.17}$$

- Spectral flux discrete derivative:

$$\Delta SF_t = SF_t - SF_{t-1}, \tag{5.18}$$

where SF_t is the spectral flux of frame t and $SFt-1$ is the spectral flux of frame $t-1$.

Following the computation of the frequency domain features, basic descriptive statistics, including the mean, maximum, minimum, standard deviation and root mean squared are calculated. This leads to a 50-dimensional feature vector (10 features x 5 descriptive statistics), as presented in Table 5.2.

5.4.2.3 Mel-based Features

MFFCs are features commonly used in the processing of speech and audio that are derived from the power spectrum of a signal and aim to represent its spectral characteristics. They are often used in speech recognition systems, speaker identification systems, and music genre classification systems. MFCCs can help extract and classify important features from a signal due to their ability to concisely capture the spectral characteristics of the signal. They can be used to

Table 5.2: The 50-dimensional feature vector of the frequency domain features, represented as a table (10 features x 5 descriptive statistics). The details of the vector are presented as descriptive values of the feature set.

Feature	Mean	Max	Min	STD	RMS
BER	mean_ber	max_ber	min_ber	std_ber	rms_ber
delta BER	mean_delta_ber	max_delta_ber	min_delta_ber	std_delta_ber	rms_delta_ber
SC	mean_sc	max_sc	min_sc	std_sc	rms_sc
delta SC	mean_delta_sc	max_delta_sc	min_delta_sc	std_delta_sc	rms_delta_sc
SBW	mean_sbw	max_sbw	min_sbw	std_sbw	rms_sbw
delta SBW	mean_delta_sbw	max_delta_sbw	min_delta_sbw	std_delta_sbw	rms_delta_sbw
SRO	mean_sro	max_sro	min_sro	std_sro	rms_sro
delta SRO	mean_delta_sro	max_delta_sro	min_delta_sro	std_delta_sro	rms_delta_sro
SF	mean_sf	max_sf	min_sf	std_sf	rms_sf
delta SF	mean_delta_sf	max_delta_sf	min_delta_sf	std_delta_sf	rms_delta_sf

represent the spectral characteristics of a signal in a compact form, making them useful for feature extraction and classification tasks.

The process of implementing MFCC follows a standard method of implementation, the first stage of this process is the pre-emphasis stage. Pre-emphasis is used to boost the frequency components of the audio signal to compensate for potential loss that may occur when recording or processing audio segments. A first-order high-pass filtering method can be applied through a backward difference [243, 244] to improve the intelligibility of speech recordings:

$$y[n] = x[n] - \alpha x[n-1], \tag{5.19}$$

where $x[n]$ is the input signal, and α is the pre-emphasis coefficient.

Frame blocking is then used to divide the audio into overlapping frames. A windowing function is then applied to each of these frames to limit the chance of spectral problems. The windowing function can therefore be applied, which uses the Hamming window formula [244, 245] presented below:

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N}\right), \quad 0 \leq n \leq N-1, \tag{5.20}$$

where n is the sample index and N is the window size.

The MFCCs is then calculated by taking the power spectrum of a Fast Fourier Transform of the previously windowed signal and then applying a Mel-scale filter bank to the resulting spectrum. The Mel-scale is a nonlinear scale that is based on the perceived frequency response of the human auditory system. The filter bank consists of a set of triangular filters spaced at equal intervals on the Mel-scale. The equation for a Fast Fourier transform [246] is as follows and provides the magnitude of the signal at the defined frequency:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N}, \quad k = 0,\ 1,\ ...,\ N-1, \tag{5.21}$$

where N is the length of the signal.

The equation for Mel-scale filtering [244] which is a perceptual scale that models how humans hear frequency is defined as:

$$\text{mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \tag{5.22}$$

where f is the frequency of the signal in Hz.

The filter output is then transformed using the discrete cosine transform (DCT) [244, 247], resulting in a set of coefficients that capture the spectral characteristics of the signal in a compact form. The resulting output is the coefficients which represent the MFCCs.

$$C_n = \sum_{m=0}^{M-1} M_m \cos\left(\frac{\pi(n+0.5)m}{M}\right), \tag{5.23}$$

where M is the number of filters and n is the coefficient index.

5.4.3 Model

The model for the audio modality is based on the extracted features, including the time domain, frequency domain, and MFCC. The model is basic in nature due to the expectation that the algorithm will support the fusion approach. The MFCC features are fed through three layers of a 2D CNN, with the time-domain features being fed through a dense NN. In search of the optimal model, the research

explored the following hyperparameter spaces: number of hidden layer (0, 1, 2, 3), dropout rate (0, 0.1, 0.5), number of hidden nodes (32, 64, 128), activation function ('ReLU', 'linear'), Adam optimiser with learning rate (6.25e-3, 6.25e-4, 6.25e-5, 6.25e-6), and number of epochs (1, 5, 10). To reduce over-fitting, a Regularization lambda of λ=0.01 is utilised and batch normalization after every layer. The parameters of the final selected three CNN layers are as follows:

$$\begin{aligned} L_1 &= (K : [3,3], S : (2,2), D = 0.5, F = 32) \\ L_2 &= (K : [3,3], S : (2,2), D = 0.5, F = 32) \\ L_3 &= (K : [3,3], S : (2,2), D = 0.5, F = 64) \end{aligned} \tag{5.24}$$

where, L_i is the layer ID, K is the Kernel, S is the Stride , D is the Dropout and F is the Filter size.

Based on the test, the remaining model parameters and optimisations could be selected. Parameter settings were selected based on the collected results, 3 layers were chosen, with 128 hidden nodes each, the dropout value was set to 0.5, and each model is trained using the Adam optimiser (with a learning rate of 6.25e-4) and a batch size of 32 for 1 epoch. The activation function for both datasets in the model was selected to be ReLU. ReLU was selected because it was computationally efficient for the edge processing task; however, awareness was needed of potential issues with zero-output neurones. The ReLU activation function can be written as:

$$g(z) = max(0, z) \tag{5.25}$$

where z is the input of the activation function.

A DNN was used for the time domain features, which is a type of network where every neuron in a layer is connected to every neuron in the following layer. The ReLU activation function is also included to introduce non-linear results into the network, enabling the ability to learn complex relations between the input and output data. The formula for a DNN is as follows:

$$
\begin{aligned}
Z^{[1]} &= X \\
Z^{[l]} &= g^{[l]}(W^{[l]}Z^{[l-1]} + b^{[l]}) \qquad \text{for } 1 < l < L \\
Y &= g^{[L]}(W^{[L]}Z^{[L-1]} + b^{[L]})
\end{aligned} \tag{5.26}
$$

where $Z[l]$ is the activation vector for layer l, and $Z[0]$ is equal to X.

An F1 score was selected as the basis for the measurements, allowing the resulting values to be reported against other articles and research. The F1 score also allows a single method of comparison to be presented, allowing cross-modality comparisons to be performed when the existing literature does not yet exist.

5.4.4 Results

Table 5.3: Results displaying the values of the time-frequency domain features when combined with the MFCC and time-frequency results. The methods are compared using the F1 score.

Classification Method	**F1 Score**
Time-Frequency Domain + Support Vector Machines	0.6478
MFCC + Time-Frequency Domain	0.8032

The fusion of the extracted audio features resulted in F1 = 0.80, as presented in Table 5.3. The combination of MFCC and time-frequency features provided a method of effectively combining the features and produced a relatively high F1 score. The fusion of audio features is useful in the context of audio data sources, which can then be further improved by fusion with the transcription text results. The F1 score for the audio modality is successful compared to the results reported in a similar literature; for example, an F1 score of 0.67 has previously been reported in the context of the detection of depression using audio modalities [248]. This has potential to be further improved through a more in-depth dataset being curated, to enable a larger corpus of audio-based data to be processed. Compared to baseline measure, the audio techniques performed well.

The low score of the SVM approach indicates the importance of the additional context provided in MFCCs as a processing method for violent conversation de-

tection. The initial measure conducted during the research was SVM, which is reported in Table 5.3. This model presented a low F1 score of 0.65 for the resulting algorithm, highlighting the effectiveness of MFCC + time-frequency domain model. The contrasting results in this table showcases the importance of MFCC features being included in the model, where the addition of frequency features improves the potential detection of violent conversation.

5.4.5 Audio Diarisation

Audio diarisation is the process of dividing audio recordings into several speech segments to then identify the number of speakers. In this work, audio diarisation was used to determine if this could be determined from the collected dataset. A research artefact was developed that extracts the number of speakers from a set audio segment. To test the usefulness of audio diarisation through speaker diarisation, a rapid prototype of a method was produced. The prototype was developed using PyTorch and was based on the pyannote.audio [249, 250] library in version 1.1. Using pyannote.audio, it was possible to examine the effectiveness of a pre-trained model in extracting the number of speakers from a set audio recording.

The prototype did not achieve positive results, and the pre-trained model incorrectly analysed every example conversation tested from the dataset described in Chapter 3. Some example graphs of these results are presented in Figure 5.3. The resulting metrics from the pyannote.metrics library were also not positive, with a high error rate of 81% reported. These results were not positive due to the lack of an accurate training dataset with a focus on violent multi-person conversations. Another potential limitation in this regard is the lack of accurate ground-truth testing for the provided audio segments. A final limitation could be the small segment size, often meaning that a total of two individuals would be speaking at one time due to the nature of conversations in the dataset. While a method of including diarisation could be useful, it was determined to be beyond the scope of the work in this project.

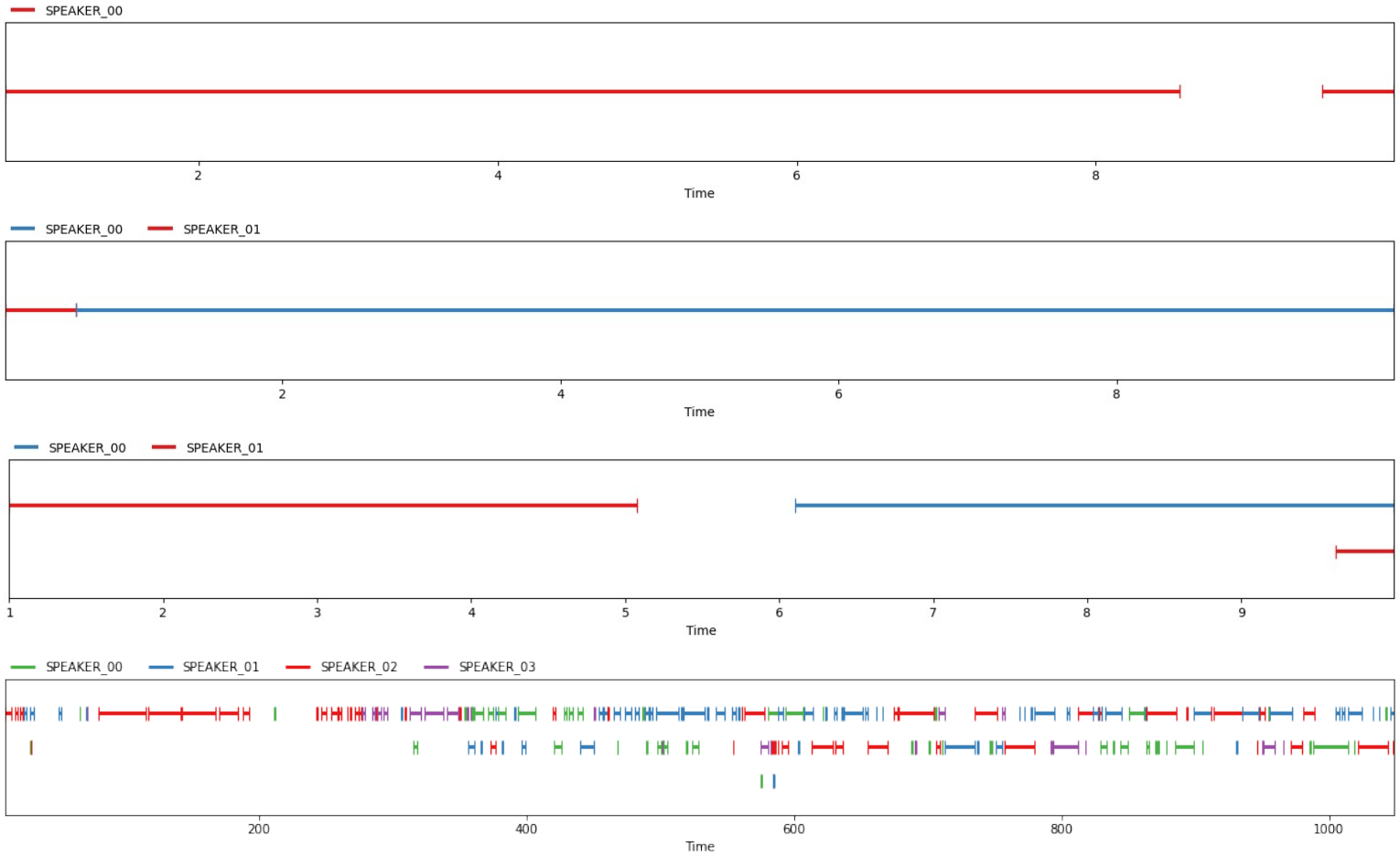

Figure 5.3: Two examples of audio diarisation results classified using a pre-trained model for detecting the number of speakers in the conversational segments. The first examples (first three) present 10-second segments, while the fourth example (last) presents a longer recording of two speakers with background noise classified as four speakers with occasional overlaps.

5.5 Fusion for Mulitmodal Data

The multimodal approach has been used for emotion recognition using audio, video, and physiological data [236]. Other researchers have incorporated the fusion of audio, visual and text information [251,251–254] during which the main focus was to advocate a multimodal approach over a unimodal approach. The section of the chapter will focus on the multimodal approach taken in this project, using the fusion of BERT, LIWC, MFCC, and time-frequency features.

5.5.1 Experimental Setup

The experimental setup used for this stage follows the methodology presented in 3. The aim is to design and develop the fusion model architecture, as well as to test the accuracy with the baseline model and other fusion attempts. For the implementation of the work, a 2017 MacBook Pro with a 2.5 GHz dual core Intel Core i7 was used for prep-rocessing of the data. The programming was primarily carried out on Google Colab which provides a single 12GB NVIDIA Tesla K80 GPU that can be used for up to 12 hours continuously. In addition, the following libraries were utilised:

- JSON: used for working with JSON data,
- NumPy: used for numerical operations on arrays and matrices
- TensorFlow.keras: used for building deep learning models
- Matplotlib.pyplot and Seaborn: used for creating visualisations of the data
- Pandas: used for data manipulation and analysis
- Math: used for mathematical operations
- Sklearn.preprocessing: used for pre-processing the data before training the models
- Warnings: for issuing warning messages during the programming process

5.5.2 Multimodal Fusion Process

The multimodal fusion process includes the identified audio and text modalities to enable a combined model to be used. The approach taken as part of this work is a feature-level fusion, where text and audio modality data is taken at a low-level feature representation and used to enable information to be processed effectively. The goal of multimodal fusion is to extract complementary information from different modalities and improve the overall performance of the system. Multimodal fusion has several ways to integrate information from different modalities, these methods could include:

5.5.2.1 Early Fusion

Early fusion is included early in the machine learning pipeline. Early fusion vectors are included and combined into a single vector representation before being included and fed into the model. This approach is similar to the chosen fusion approach, however, differences relate to the early fusion approach occurring before any processing or feature extraction is performed on the final vector. The formula for an early fusion approach could be the following:

$$X = [T; A], \tag{5.27}$$

where T and A represent audio and text features respectively, X is the output fusion vector, and the ";" represents the concatenation of the data.

5.5.2.2 Late fusion

Late fusion is the opposite of early fusion, where features are not fused together until very late in the process. In late fusion, the predictions made by models trained on individual modalities are combined after the models have made predictions; this is usually achieved through averaging the results. The formula for late fusion is as follows:

$$X = wT * T + wA * A, \tag{5.28}$$

where T and A represent the audio and text features respectively, X is the output prediction, and w represents the weights assigned to the audio and text features.

5.5.2.3 Feature-level Fusion

As an alternative to early and late fusion, feature-level fusion is where the outputs of models on trained modalities are combined at the feature-level. In some cases, this can be performed through normalisation, transformation, reduction schemes, or any other form of concatenation approach. The formula for feature-level fusion is as follows:

$$X = f(T, A), \tag{5.29}$$

where T and A represent the audio and text features respectively, X is the output prediction, and f represents the concatenation of the selected features.

5.5.2.4 Element-wise Fusion

Element-wide multiplication integrates feature vectors from multiple modalities by outputting the element-wise product of each vector element. Therefore, the resulting vector element includes the previously identified modalities of the same length as the input modalities. By multiplying the elements of the feature vectors, it is possible to capture the interactions between features of different modalities. The formula for element-wise multiplication is as follows:

$$X = T \odot A, \tag{5.30}$$

where T and A represent the audio and text features respectively, X is the output prediction, and $\odot$ represents the element-wise product.

When the element-wise multiplication is applied to the specific context of this task, the formula can be expanded to be written as follows:

$$X = TbTl \odot AtdAm, \tag{5.31}$$

where T and A represent the audio and text features respectively, X is the output prediction, b represents BERT, l represents LIWC, td represents time domain, m

represents MFCC, and $\odot$ represents the element-wise product.

5.5.2.5 Selected Technique

In the case of this research, early fusion could not be applied to the large nature of the dataset, meaning that an early fusion would combine data too early in the process for meaningful results to be formed. Late fusion was not applied due to the need for some post fully connected and softmax layers which need to be applied to the model to improve accuracy. Finally, element-wise fusion could not be applied due to the complex nature of the modalities, meaning that the modality layers were not the same size during input or output.

Therefore, these differences must be considered when selecting a multimodal fusion process. Due to the nature of the work conducted in this research, a feature-level fusion was applied due to the relevant weightings of the input modalities. In the case of the presented model, a feature-level approach of equal rankings is applied for each model, after which additional processing is performed. This extra processing means that early fusion could not be selected, while also highlighting the need for a weighting method to be included for the output algorithm. Therefore, the output of the model is based on this feature-level approach, which means that the algorithm is processed before the resulting fusion is applied, which was chosen to enable a more informed classification.

5.5.3 Model

Upon determining the modalities and the completion of the initial model for each modality, it was necessary to design and develop the overall fusion model. Figure 5.4 presents the overall framework for the proposed fusion model. Integration of multimodal fusion is achieved through the combination of four models. The concatenation of features and embeddings includes (1) BERT, (2) LIWC applied in a Bi-LSTM, (3) Time-Frequency domain, and (4) MFCC applied through a CNN. The concatenated overall model is then fed into a three-layer Fully Connected (FC) network, followed by a Softmax layer which assesses the type of label using feature-level fusion. The audio and text concatenation captures short-term, as well as long-term acoustic and linguistic characteristics to detect violence level

in conversations. The softmax layer enables a classification on the model to be determined using binary classification, where 0 is non-violent conversation segments, and 1 is violent conversation ranking.

To fuse the four types of information extracted from the different modalities, embeddings generated from both the BERT and the Bi-LSTM model along with the 2D MFCC CNN representations and the Audio Time Domain Dense layer are integrated. The concatenated embeddings are then passed to three-layer FC networks, which serves as a merge step. The concatenation of the embeddings is defined in the following equation:

$$\begin{aligned} a &= BiLSTM(X_{embeddings}), \\ b &= TimeDomainF(X_t), \\ c &= CNN2D(X_{rep}), \\ d &= BERT(X_{embeddings}), \\ x_{fuse} &= [a,\ b,\ c,\ d], \end{aligned} \tag{5.32}$$

where X represents the input modality data from previous processing.

Given that the extracted features exhibit differing formats, different models are proposed regarding the resultant structure of the different feature sets. The approach used in this model attempts to use a logicial format of each individual modality first running, being concatenated, and then the resulting output value being classified. The first considered in this scenario is MFCC, which, as previously described, is presented through three 2D CNN layers before being included in the concatenation process. The time domain features, which represent both the time and frequency feature extraction, are fed through a DNN to preserve discriminative information, before the output of this is used in the concatenation process. The input and output shapes for these modalities are presented in Figure 5.4, with representations of the overall task also included. At this stage, the audio modalities are ready for concatenation.

For the text modalities, an extra set of pre-processing is included, to remove any extra details relating to the text-based content. For the LIWC data, a Bi-LSTM is used with an attention layer. For this process to be effective on the edge, the previously identified PCA is completed for dimensionality reduction.

The post-processed transcribed text is then also used in the BERT model, the output vectors of which are included in the concatenation process. No further processing is applied to BERT, due to the already large set of vectors which this process generates.

The concatenation is then applied through feature-level fusion, with each model having an equal vote based on the output result, which is then processed through three more FC layers to ensure an accurate output classification can be generated and is a common use case in deep learning architectures. The benefits of applying FC layers to the output of feature-level fusion is four-fold, this includes: (1) non-linearity can assist in capturing more complex patterns from data, (2) new features can be learned that are a combination of features from the input modalities, (3) the model can become more robust and noise from previous features can be reduced, and (4) the scalability of the model can be improved due to the high-dimensionality of the feature-level fusion. Therefore, it was necessary for three FC layers to be applied after concatenation. However, potential risks of overfitting mean that it is necessary to carefully tune the parameters of the model to enable good and generalised performance. For the FC layer, a basic formula can be used:

$$FC = ReLu(Wx + b) \tag{5.33}$$

where x is the input of the concatenation task, W is the weight of the matrix size, b is the bias vector size, FC is the output and ReLu is the activation function.

The best performing model after testing consisted of 3 hidden layers, with 128 hidden units in each layer. The final layers were selected to be: FC1=128, FC2=128, and FC3=128. When considering that the FC layers are stacked, this formula can be applied three times in a recursive sequence, where each output layer is the input to the following layer. Therefore, the formula for the FC layers is as follows:

$$\begin{aligned} FC1 &= ReLu(W1x + b1), \\ FC2 &= ReLu(W2y1 + b2), \\ FC3 &= ReLu(W3y2 + b3), \end{aligned} \tag{5.34}$$

where x is the input from the concatenation task, W is the weight of the matrix size, b is the bias vector size, FC is the output of each layer and ReLu is the activation function.

The final aspect of the model, as represented in the framework in Figure 5.4, is the softmax function, which reduces the complex dimensionality of the model to a single output ranked 0-1. The output represents the probability of a high classification, and therefore in the context of this work can be considered to probably be of violent conversation segments. The following formula can be used as a softmax function for binary classification:

$$P(y = 1 \mid x) = \frac{e^{z_1}}{e^{z_0} + e^{z_1}}, \tag{5.35}$$

where z_0 is the score for the negative class, z_1 for the positive class, P is the probability output function, x in the input value, and e is the exponential function.

The model loaded data from a number of models, to ensure each imported shape was correctly imported, each shape was checked. The X, Y training, validation, and testing values for each of the output shapes for each of the models is presented in the following code output. The output should be noted that the Y values relate to the headers of the columns and are therefore single rows of outputs. The code output below presents a complete overview of this shapes, and has been checked to ensure each training/testing split has been correctly applied and each model correctly imported.

```
MFCC and Features: (1028, 440, 26) (1028, 145)
LIWC: (1295, 35)
BERT: (30520, 768)
X_train_mfcc shape:  (822, 440, 26, 1)
X_val_mfcc shape:  (200, 440, 26, 1)
X_test_mfcc shape:  (6, 440, 26, 1)
X_train_feat shape:  (822, 145)
X_train_liwc shape:  (822, 35)
X_train_bert shape:  (822, 768)
X_val_feat shape:  (200, 145)
X_val_liwc shape:  (200, 35)
```

```
X_val_liwc shape:  (200, 768)
X_test_feat shape:  (6, 145)
X_test_liwc shape:  (6, 35)
X_test_liwc shape:  (6, 768)
y_train_mfcc shape:  (822,)
y_val_mfcc shape:  (200,)
y_test_mfcc shape:  (6,)
y_train_feat shape:  (822,)
y_train_liwc shape:  (822,)
y_val_feat shape:  (200,)
y_test_feat shape:  (6,)
y_val_liwc shape:  (200,)
y_test_liwc shape:  (6,)
```

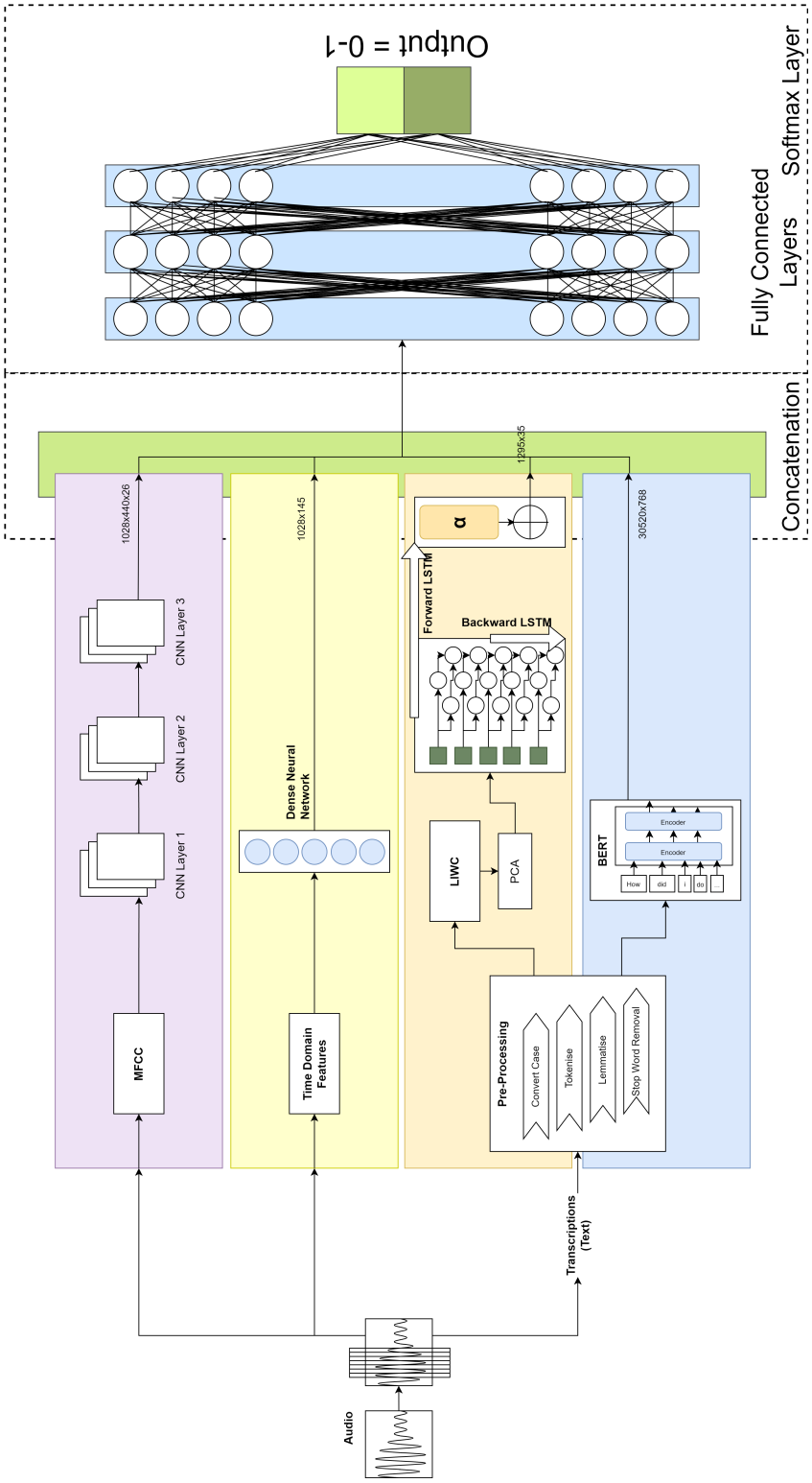

Figure 5.4: A complete model of the fusion data model, from the initial audio segments to the final binary classification output. The multimodal fusion approach uses four main processes (presented from top to bottom): MFCC, time-frequency domain features, LIWC, and BERT.

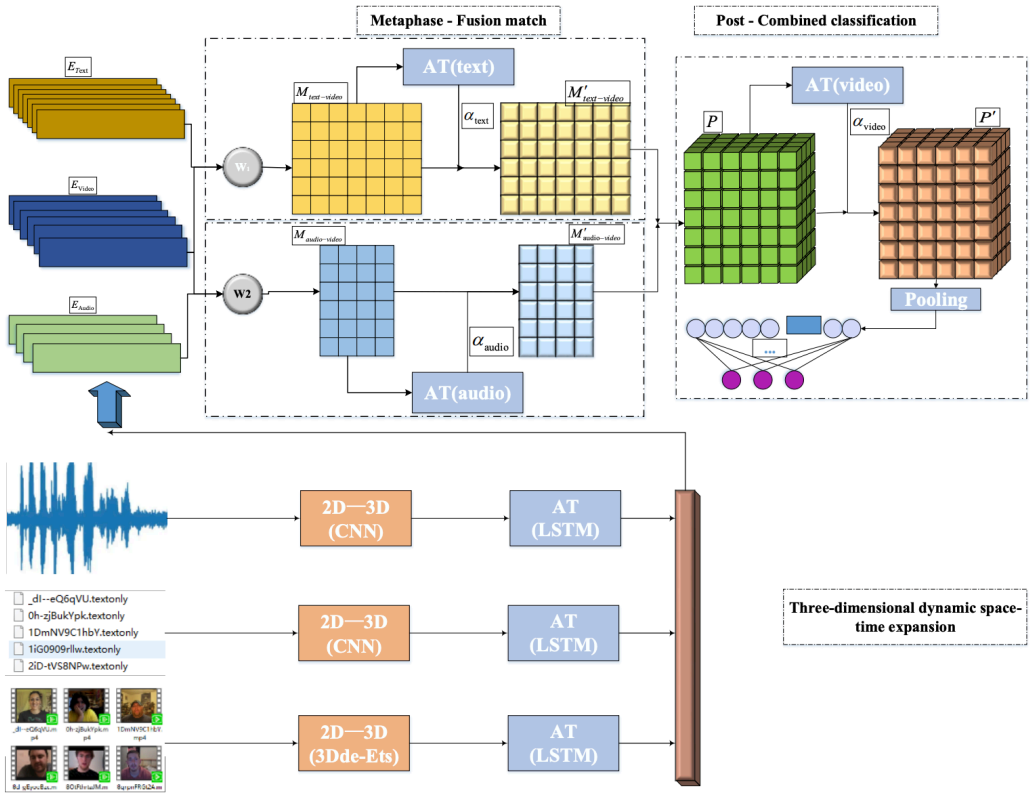

Figure 5.5: A diagram presenting a comparative architecture for the purposes of emotion classification using multimodal fusion, as presented in existing literature (diagram source: [2]).

To ensure the novelty of the architecture proposed in this thesis (presented in Figure 5.4), similar and relevant literature was explored. The alternative models reviewed are both published in 2024 and highlight the current approaches in similar contextual areas. The first comparison architecture is presented in Figure 5.5 and presents an emotion classification system based on multimodal fusion [2], with similarities to the work conducted in this thesis on violent language detection. The presented architecture [2] introduces a multimodal fusion approach that emphasises spatial and temporal feature enhancement to address dynamic correlations both within and across different modes. The architecture [2] contrasts to the method proposed in this thesis by combining three modalities including audio, text, and video footage. The approach [2] aims to capture and model both short-term and long-term dynamic interactions between various modes, leverag-

ing the strengths of the proposed framework [2]. The approach is different to the method presented in this thesis, with no additional context through methods such as LIWC or similar. Instead the authors are interested in a spatial-temporal model, using 2D-3D CNNs and LSTMs for each modality before fusion matching. The work proposed in the new architecture attempts to focus on more contextual algorithms such as BERT, LIWC, MFCC, and time-frequency domain features.

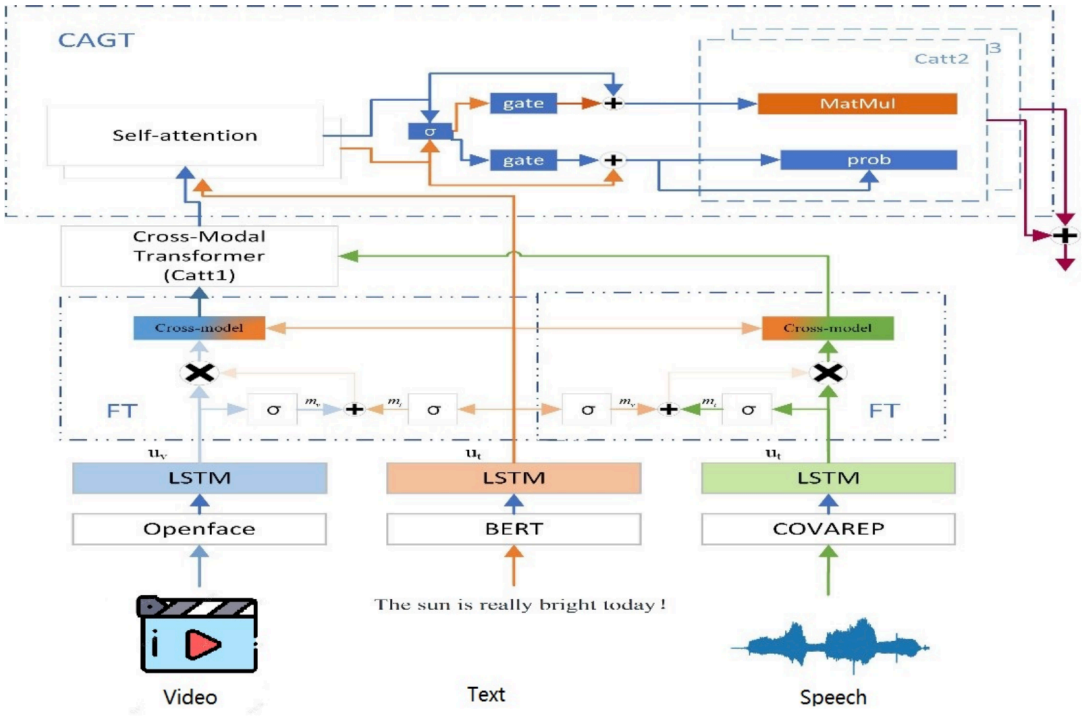

Figure 5.6: A diagram presenteing a second comparative architecture for the purposes of emotion analysis using three multimodal data sources (diagram source: [3]).

A second example of a multimodal emotion classification architecture is presented as a diagram in Figure 5.6. The work [3] proposes a network employing a causal gate attention mechanism for cross-modal fusion was introduced to improve harmony and complementary across the modalities of video, text, and audio. The model makes use of Openface for video, BERT for text, and COVAREP for speech, each is then used with a LSTM. In contrast to the architecture used in this thesis, both presented in existing literature [2,3] focus on single algorithms

for each modality, while also incorporating video-based data. The approach used by the authors [3] enables the model to enhance interactions between modalities through iterative processes, effectively bridging audio-visual bimodal data with textual unimodal representations. The work conducted in this thesis instead focuses on adding context to two data formats, which is necessary due to the expectation of the algorithm working in domestic settings or while on the edge, where extra context from existing data becomes more important due to data recording constraints.

5.5.4 Results

Table 5.4: Results displaying the values of the multimodal fusion approach including MFCC, time-frequency domain features, BERT, and LIWC combinations. The methods are compared using the F1 score.

Model	**F1 Score**
MFCC + Time-Frequency Domain + BERT	0.7790
MFCC + Time-Frequency Domain + LIWC	0.6934
MFCC + Time-Frequency Domain + LIWC + BERT	**0.8454**

This section presents the results of the proposed techniques with the combination of both the text and audio features and the methods already used in the analysis. Combining MFCC, time-frequency domain, and BERT resulted in an F1 score of 0.78, which is an increase of the text-based fusion reported in Chapter 4, however, lower than the audio fusion technique when implemented as an individual modality. A similar case can be seen with the fusion of MFCC, time-frequency domain, and LIWC, which resulted in an F1 score of 0.69, which is also lower than the MFCC and time-frequency domain feature result, indicating the previously identified issues with the text-based fusion technique potentially requiring a larger corpus of transcriptions. Finally, the results of the complete fusion model that includes MFCC, time-frequency domain, LWIC, and BERT is F1 = 0.85, indicating an improvement in the individual modality methods for both text and audio. The higher F1 score for the full fusion approach indicates

that the multimodal features are effective at supporting other modalities, especially considering the relatively low F1 scores for text as an individual modality that were reported.

The lack of existing literature on the detection of violent language makes it difficult to compare results against a baseline; however, the resulting F1 scores provide an impetus for ongoing work in multimodal, multi-level fusion techniques. The baseline model was therefore determined to be a random forest of all features which achieved an F1 of 0.75, significantly lower than the reported results for the MFCC, time-frequency domain, LIWC and BERT fusion model. The improvement of the model over each individual modality and fusion approach indicates how the resulting model is more effective at detecting violent language by considering a larger selection of both audio and text data. The model indicates the importance of a careful selection and combination of methods which provide additional context (e.g., the combination of BERT vectors and LIWC analysis).

Based on the small dataset size, the resulting model of MFCC, time-frequency domain, BERT and LIWC features was found to be effective. Despite this, the results should be noted considering that the F1 score of 0.85 means that 15% of the segments were not correctly identified, which is significant in the context of detecting violent language, which requires more work to examine the effectiveness of the dataset, before the work could be included in applied situations. While the output model size is capable of being placed on an edge device, further considerations such as reducing the dimensionality of MFCC and BERT may be required.

5.6 Discussion

Table 5.5: The full results reported throughout this thesis, containing the baseline methods, text approach, audio approach, and the final multimodal fusion approach. The methods are compared using the F1 score.

Model	**F1 Score**
All: Baseline Model (Random Forest)	0.7469
Text: CNN + Glove	0.6371
Text: NN + Glove	0.6602
Text: LSTM + Glove	0.5946
Text: LIWC + K-Nearest Neighbors	0.5964
Text: LIWC + Random Forest Classifier	0.5913
Text: BERT + LSTM	0.6602
Text: BERT + CNN	0.6681
Audio: Time-Frequency Domain + Support Vector Machines	0.6478
Audio: MFCC + Time-Frequency Domain	0.8032
Fusion: MFCC + Time-Frequency Domain + BERT	0.7790
Fusion: MFCC + Time-Frequency Domain + LIWC	0.6934
Fusion: MFCC + Time-Frequency Domain + LIWC + BERT	**0.8454**

This chapter introduced the audio and fusion aspects of the research, within this discussion the results of these techniques will be further discussed in the context of the application area and in comparison to other reported results. Notably the results will then be discussed in the context of the overall aim of the project, and the potential application area of the work to be implemented on edge technologies such as a mobile phone or smart watch. The techniques reported in this chapter were overall considered a success, due to the positive demonstration of the system and the effective results achieved; however, some further findings of these techniques present novel applications and limitations for future work and therefore should be considered by future researchers.

The results of the MFCC and time-frequency domain feature extraction method were positive, with audio features achieving a positive F1 score of 0.80, which

when compared to existing F1 scores in a similar contextual area is high. For example, a paper on the detection of depression through vocal, facial, and semantic cues had a F1 score of 0.50 [227]. A slightly higher approach was found in audio features for detecting depression through multimodal data, in which a F1 score of 0.77 was achieved [226]. The result also performed well in comparison to the baseline model for fusion reported in Table 5.5, with a baseline F1 score of 0.75. While the audio results yielded positive results, it was prevalent that the language and textual context of the spoken language was not being analysed effectively, meaning that the features collected were effective for statistical and acoustic analysis, but not useful in the context of the language being used, or the words spoken. The audio results were overall positive and provided a positive impetus for using the technique, despite more novel models existing, the proposed methods found successful improvement on the baseline model and the previous results.

Using the audio modalities identified in this chapter and the text NLP used in Chapter 4, it was therefore possible to fusion the resulting modalities from the initial dataset. The fusion approach yielded varying results, based upon the fusion methods used and the modalities and processing applied, the results of which are presented in Table 5.5. The results of the MFCC and time-frequency domain features was higher than when BERT and LIWC were used individually (F1 = 0.78 and 0.69 respectively), however when applying both textual models with both audio methods a higher F1 score of 0.85 was achieved. This may be due to the models and processing being chosen to effectively support the other modalities and methods used. For example, the MFCC and time-frequency domain features compliment the analysis methods of each method, and the same applies to the lingusitic approach of LIWC and the comprehensive methods used in the BERT model. The results, while positive, should be considered in context. For example, while a F1 score of 0.85 is high, it still leaves 15% of 10-second segments as an inaccurate classification, which should be the aim of improvement for future work. However, the 0.85 F1 score for the overall model was a significant improvement on most previous work, and a similar F1 score to the model presented in [248], however, the dataset used for training and testing the model presented in this chapter was much smaller, and therefore future work

could consider the comparison of this approach with other datasets.

The decision to prioritise F1 scores over precision and recall in classifying audio as violent or non-violent stems from the critical need to carefully balance the accuracy of predictions against the risks of mislabeling. High precision ensures the precise identification of violent content, minimising the chance of falsely tagging non-violent audio as violent, which is crucial to avoid unnecessary alarm or action. Conversely, focusing on high recall aims to capture as many violent instances as possible, but comes with the risk of higher false positives, which could dilute the seriousness of the actual violent content or misguide subsequent interventions. In contexts where accurately identifying violence is paramount to ensure appropriate responses and interventions, mislabeling can have significant consequences. Therefore, the F1 score, which provides a harmonic mean of precision and recall, was selected as the optimal metric to ensure that the classification is accurate and reliable, minimising the potential for harmful mislabeling in sensitive scenarios.

The purpose of the fusion model presented in this chapter is to detect violent conversations for the purposes of crime prevention, it is therefore necessary to discuss the work within this area. The model provides a novel approach to classifying violent conversations from a multimodal dataset, combining the modalities of text and audio through MFCC, time-frequency domain features, BERT and LIWC. The positive results of the methods presented in this chapter imply that this approach is feasible and achieves good results when the correct methods are used. However the model, while accurate, could be improved through considering a larger dataset in the future and applying a wider range of contextual attributes into the data source (e.g., how many people are speaking for the purposes of diarisation). However, the consideration throughout the project should ensure that the dataset and models used can still be transferred to edge devices. The results provide a positive impact on the area of violent conversation detection, and suggest a continued approach to using multimodal datasets and fusion models in the prevention of crime. Considerations will need to be made to the cost of such systems, requiring a consideration of the edge approach taken in Chapter 6. The next steps for the applied area will be to compare the model to other models, based on the use and fine-tuning of the model to other datasets and pro-

cesses, such as depression or mood detection as opposed to just being applied to a dataset of violent conversations.

5.7 Conclusion

This chapter presented the results of audio and fusion techniques for detecting violent conversations in the context of crime prevention. The audio results reported in this chapter report a positive F1 score of 0.80, indicating the extraction and processing of MFCC and statistical features is an effective approach at classifying violent conversations from audio segments. The audio results provide a positive method of detecting data from audio, beyond that of the specific contextual language used. The fusion approach achieved an effective F1 score OF 0.85, also indicating that the LIWC and BERT research conducted in the previous chapter can effectively improve the results of the MFCC and time-frequency domain features. The results of the fusion indicate that the multiple modalities can provide an effective, but complex method of detecting violent conversations. In addition to the audio and fusion methods, the chapter reported on audio diarisation, which due to limitations of the dataset was not found to be an effective method of extracting details from the audio segments. The results in the chapter provide a positive impetus for future work focused on multimodal data for crime prevention, specifically in audio and text literature. The next aim of the work is to enable a method of running the complex multimodal model on the edge, which is a computationally complex challenge in the context of crime prevention.

Chapter 6

Real-Time Processing on the Edge

6.1 Chapter Overview

This chapter presents the real-time processing on the edge of violent language detection for the purposes of preventing domestic violence. Edge computing refers to running the processing as close to the periphery of a network as possible. The edge in this scenario refers to the edge device of the user, as opposed to an edge node [255]. The work in this chapter presents the model developed during this project in user-based computing scenarios, which includes a simulated smart home device and a mobile application for Android and iOS.

There has been a growing interest in using edge computing technologies to prevent crime [256]. Faster response times, lower bandwidth needs, and enhanced data security are just a few potential benefits of using edge technologies [257] that could support the prevention of crime. Law enforcement organisations and other interested parties can learn a lot about criminal activities and take preventive action using distributed computing to analyse and react to data from various sources, including CCTV [46], sensors [258], and smart home hubs [259]. Crime prevention on the edge is a developing field that could change how crime is prevented in real-time.

This chapter provides an overview of the edge computing system and details

the overall context of the work. The chapter is split into the following sections: first an introduction to the work is presented, followed by design considerations for the work performed including smart home and mobile considerations. The next section introduces the implementation of the system, which focusses on the setup, development, and demonstration of edge processing applications for smart homes and mobile devices. The resulting device is reviewed, features are presented, and future considerations are identified as part of the results. Finally, a discussion on the challenges and opportunities presented by edge computing in the scenario of the study is presented, followed by the conclusion and identification of future work.

6.2 Background

Real-time processing on the edge refers to the ability to process data and make decisions close to the data source. This can be contrasted with traditional approaches in which data is collected, transmitted to a central location, and then processed. This enables data processing and decision making to occur in real time using local computing resources [257], rather than relying on a remote server or the cloud. Allowing faster response times and less dependence on connectivity, making it suitable for applications that require immediate action or those with an unreliable connection to the cloud.

Real-time processing can greatly improve crime prevention efforts by enabling devices to quickly identify and respond to potential threats such as those identified by crowd sourcing [260]. For example, a mobile device equipped with a camera could recognise faces through real-time processing [261]. Despite this, most of the work is not able to compute this analysis on-device or on the edge of a network. Furthermore, previous reviews of work on mobile crime prevention applications identify the lack of evidence in some application designs [262]. In general, the use of edge-based real-time processing or mobile applications has the potential to significantly improve the effectiveness of crime prevention efforts. There are several benefits to real-time processing on the edge:

- Low latency: When processing data close to the source, the latency (i.e.,

the time it takes for the data to be processed and a decision to be made) could be reduced [263, 264]. This is important in applications where timely decisions are critical, such as in self-driving cars or industrial automation.

- Increased reliability: By reducing the distance that data must travel, the risk of data loss or corruption due to transmission errors could be reduced [263].
- Reduced bandwidth requirements: When processing data, the amount of data that need to be transmitted to a central location is reduced, which can help reduce bandwidth requirements [257].
- Improved privacy: By processing data on the edge of a network, it is possible to reduce the amount of personal or sensitive data that needs to be transmitted to a central location, which can help improve privacy [257, 263].
- Real-time processing on the edge can be enabled by distributed computing technologies, which allow data to be processed at the distributed end of a network, closer to the source of the data [257, 264]. This can be done by using specialised computing devices or by leveraging the processing power of devices such as smart cameras or smartphones.

6.3 Design Considerations

The purpose of the mobile application was to provide a user-centred process to identify and detect violent language on the device. Ethical concerns related to individual interviews mean that it is necessary to determine, through the existing literature and applications, the technical and design considerations that must be made for the application to work effectively. Design considerations were formed through an extensive review of the literature [155] and identification of existing application design processes.

The design considerations formed as part of this work represent the initial stages of thinking and implementation with regard to an edge processing crime prevention application. The proposed considerations were then used within the

development of the application to ensure that the application itself follows modern techniques and uses academic backing, in addition to providing the best possible method of processing the complex multimodal model on mobile devices. Most of the design considerations identified in the following sections relate to the implementation of the device in the context of the specific edge scenario; however, they can also be considered for wider applications of the work. The design considerations formed through the research team are defined as follows:

6.3.1 Self Reporting

Having the user self-report a potential issue was a key design consideration during the development of the application. As with any automated technology, precision is not perfect, and therefore users should be allowed to control recording methods and devices to ensure that reports can be made by the individual user if necessary, at any time, similar to existing applications [33, 265]. These existing applications present the opportunity for the work to consider how self-reporting could be combined with the automated reporting techniques proposed as part of the edge model. Furthermore, this could form a novel user feedback loop, which could be further used in the future to improve datasets.

6.3.2 Manual Recording

Allowing the user to manually record the audio is also a consideration linked to the previous concept of self-reporting. While existing applications [31, 33] consider manual data collection, the main focus of the application is automated data collection; therefore, it is important for designers to consider whether data can be recorded manually and how manually recorded data can be trusted. The manual recording would enable the user to feel in control of the application during use, which can increase the user confidence in the device, and provide encouragement to engage with the application to report potential crimes. Limitations of manual data would be the limited amount of data to be processed and the amount of contextual detail captured by the recording device.

6.3.3 Persistent Notification

Notifications are methods of presenting feedback to users; for more modern devices, persistent notifications could support the continuous display of information through user feedback. As the edge application requires constant data access, a persistent notification could also be shown for this purpose. A persistent notification could also provide a useful method of self-reporting and manual control to be provided to the user discretely without having to use the application, in a different approach to the shake to activate function used in previous literature [265]. The further usage of a persistent notification could also be to encourage users to recognise that the device is running and enable the edge processing to be activated at any point. Risks related to persistent notifications could include the identification of the notification to someone other than the victim, causing potential issues with the privacy of the content presented as part of the notification, or awareness of the notification itself.

6.3.4 Privacy

The nature of domestic violence and abuse means that privacy should be the primary concern when developing applications for this context. For example, existing work has considered privacy concerns about computer security interventions for survivors of intimate partner violence [25], highlighting the need for correct data handling and reliable local data storage. Despite concerns about the use of technology, the potential of technology that considers privacy has previously been explored [26], and therefore should be critical in the design of future crime prevention applications. There are several potential privacy issues that can arise when real-time processing is used to prevent crime. For example, the collection and analysis of personal data without the knowledge or consent of individuals can raise concerns about privacy and the possibility of unjust surveillance or profiling. There is also the risk that data collected for crime prevention could be misused or accessed by unauthorised parties, leading to the abuse of personal data. To mitigate these risks, organisations must have strong privacy safeguards in place and be transparent about their data collection and usage practises. This may include obtaining consent from individuals before collecting and using their

data, implementing robust security measures to protect data from unauthorised access, and clearly communicating the purposes for which data will be used.

6.3.5 Contact Warnings

Trusted contacts have been implemented in similar applications [33, 265] and are individuals the user trusts. Depending on the application, trusted contacts can link data to Police or security services, close family members, or friends. It is necessary for an application to consider the export of privacy-aware information to trusted contacts without revealing personal or compromising details. Therefore, it is a consideration for application designers to plan and design methods that can effectively share the results of the data while keeping the original files secure. There should also be considerations from designers relating to the verification of contacts, and the potential for trusted contacts to be ignored in favour of security personnel; this should therefore be considered by the designers as part of any technology design.

6.3.6 Microphone Quality

The quality of the microphone is an important consideration when designing a system that used a mobile or smart home recording device; each scenario has specific limitations for working with such data, for example, a smart home may be placed in a corner of the room, while a mobile device may be positioned too close to one individual in a group conversation. Microphone quality can depend on the quality of the onboard recording devices, such as the considerations regarding omnidirectional or unidirectional microphones, in addition to the positioning and placement of the microphone both on the device and in the wider environment (the room, location, or position of the phone on the individual's person).

6.3.7 Storage

Device storage is also a design consideration when working with the storage of data on an end-user device; initially, there is the privacy and access concern related to the removal or deletion of important recordings. In addition, there

should be considerations as to the best method of storing the initial model on the edge system, however, due to more novel technologies such as TensorFlow Lite and the removal of initial binaries and data during the process of using a pre-trained model, this is less of a problem for modern devices.

6.4 Implementation

Two implementation methods were chosen as part of the process, to select the devices, the previous literature identified in Chapter 2 was considered. For implementation purposes, the smart home device and mobile applications were chosen, due to their ubiquitous use in modern society, with a reasonable size to perform edge processing effectively. For the smart home device, a simulated smart home device in the form of a Raspberry Pi was selected. For the mobile device, development was proposed to work effectively on both Apple and Android devices. The development selection was Python for the smart home device, and Flutter/Dart for the mobile devices due to cross-compatibility and existing libraries being provided.

The Raspberry Pi smart home device being is often used in static, reasonably predictable environments; and the mobile application developed in Flutter and Dart being used in unexpected, rapidly changing environments with a potentially large pool of other people in a surrounding area. Due to this, both methods are believed to contribute to the overall study through providing multiple testing scenarios, to examine the effectiveness of the edge processing device in practice. How different scenarios work in regard to user feedback and other design considerations could also impact the model; therefore, this will require two separate development pathways to be followed.

6.4.1 Model Setup

To convert the existing TensorFlow model from the one presented in Chapter 5, a method to reduce the complexity of the model was required. The model was developed using the TensorFlow library, meaning that the processing model could be pre-trained and deployed on edge devices rapidly. Initial consideration

was taken to explore if the full model could be applied on the Raspberry Pi; however, due to poor performance, this was removed from the consideration due to inconsistency in the running of the model, and poor support from existing libraries. TensorFlow Lite enabled the work to be rapidly applied to both smart home and mobile applications, by pre-training the model and enabling a reduced size of the overall file. The benefits of TensorFlow Lite follow that of the design considerations and the overall benefits of edge devices, including reduced latency due to no server connectivity, size due to the model being pre-trained, and power consumption due to the lack of network connection.

Due to the existing code being written in Python, the TensorFlow lite model was able to be generated using existing code. The conversion code was the entire TensorFlow Lite model implementation stage, after which a file named $outputModel.tflite$ can be used to implement the pre-trained model on the edge. The complexity of the initial model can cause this to still be large; however, the output tflite file for the final model of this stage was a total of 48.3 MB, indicating the significantly smaller file size compared to the full implementation. TensorFlow Lite was used to also implement the model on the devices, with supported libraries in both Flutter and Python environments.

Once applied in the selected programming scenario, the model can be investigated to ensure that the correct input and output shapes are stored. Therefore, the output shape can be presented as an array with the shape $[1, 2]$, using a 32-bit float as the output. This is considered correct, due to the overall model implementing binary classification. Therefore, the output details of the model object can be represented as:

```
[{
    "name":"StatefulPartitionedCall:0",
    "index":48,
    "shape":"array("[1,2], dtype=int32),
    "shape_signature":"array("[-1,2], dtype=int32),
    "dtype":"<class""numpy.float32"">",
    "quantization":(0.0,
    0),
    "quantization_parameters":{
```

```
     "scales":"array("[], dtype=float32),
     "zero_points":"array("[], dtype=int32),
     "quantized_dimension":0
    },
    "sparsity_parameters":{
    }
}]
```

The multimodal nature of the model requires that the input shape of the model be more complex, with four different data types expected to be included as part of the input details. The four shapes require differing input shapes, based on the pre-processing performed in Chapter 5. The shapes should be expected to match the previous input shapes, apart from the position of the [0] array, which should instead represent the one input value. As presented in the code below, the $[1, 145]$ shape matches the expected input shape for the Features, the $[1, 35]$ shape matches the expected input for the LIWC features, the $[1, 768]$ matches the expected input shape for the BERT features, and the $[1, 440, 26, 1]$ matches the expected input shape for the MFCC extraction. The overall input details for the Lite model are presented as follows:

```
[{'name': 'serving_default_stat_features:0',
  'index': 0,
  'shape': array([  1, 145], dtype=int32),
  'shape_signature': array([ -1, 145], dtype=int32),
  'dtype': numpy.float32,
  'quantization': (0.0, 0),
  'quantization_parameters': {'scales': array([], dtype=float32),
   'zero_points': array([], dtype=int32),
   'quantized_dimension': 0},
  'sparsity_parameters': {}},
 {'name': 'serving_default_stat_liwc:0',
  'index': 1,
  'shape': array([ 1, 35], dtype=int32),
  'shape_signature': array([-1, 35], dtype=int32),
```

```
  'dtype': numpy.float32,
  'quantization': (0.0, 0),
  'quantization_parameters': {'scales': array([], dtype=float32),
   'zero_points': array([], dtype=int32),
   'quantized_dimension': 0},
  'sparsity_parameters': {}},
 {'name': 'serving_default_stat_bert:0',
  'index': 2,
  'shape': array([  1, 768], dtype=int32),
  'shape_signature': array([ -1, 768], dtype=int32),
  'dtype': numpy.float32,
  'quantization': (0.0, 0),
  'quantization_parameters': {'scales': array([], dtype=float32),
   'zero_points': array([], dtype=int32),
   'quantized_dimension': 0},
  'sparsity_parameters': {}},
 {'name': 'serving_default_mfccs:0',
  'index': 3,
  'shape': array([  1, 440,  26,   1], dtype=int32),
  'shape_signature': array([ -1, 440,  26,   1], dtype=int32),
  'dtype': numpy.float32,
  'quantization': (0.0, 0),
  'quantization_parameters': {'scales': array([], dtype=float32),
   'zero_points': array([], dtype=int32),
   'quantized_dimension': 0},
  'sparsity_parameters': {}}]
```

6.4.2 Software Loop

Figure 6.1 presents the planned software loop for the edge processing device, allowing a fully embedded method to engage with, collect, and process the results on the device. Therefore, this software loop is consistent across both the smart home and mobile platforms. Having this main software loop means that the

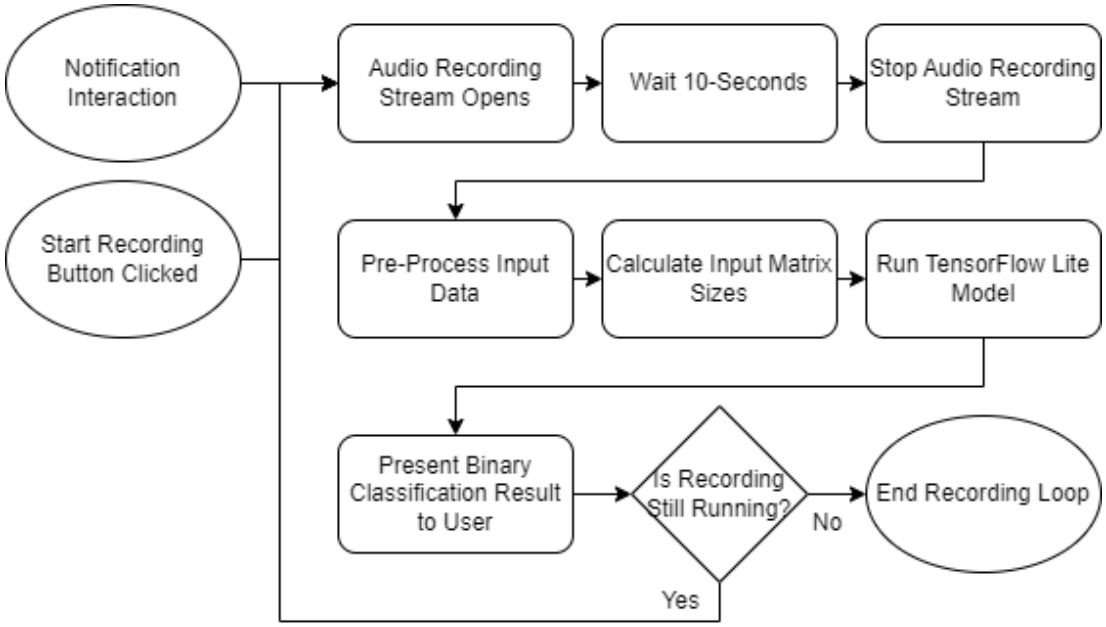

Figure 6.1: The overall software loop proposed for the edge computing devices, the loop begins with a form of interaction (notification, device turned on, button clicked) before the main loop runs.

amount of custom and specific development per platform is limited due to the overall software loop being used instead for this purpose. The software loop therefore fits with the specific development and software loops proposed for both the mobile and smart home platforms, which are presented in the remainder of this section.

The initial stage of the software loop is for the activation to begin, whether this is a notification interaction on the mobile platform, or the start recording button on either the mobile or smart home platform. Following this process, an audio stream is opened, using the provided onboard microphone or external microphone recorder; this is activated for 10 seconds before the audio stream is stopped and used for processing. This audio stream is then again opened concurrently, to ensure that no audio content is missed during the loop.

The processing stage then begins, where the input data is pre-processed; this includes saving the file to the device storage, assigning a timestamp, calculating the input features, extracting the audio features, extracting the BERT and LIWC

features, and transcribing the text for the processing. These stages largely follow similar methods as mentioned previously in Chapter 3 of the thesis. Using the pre-processed data, it is then possible to calculate the input matrices, using the previously defined input matrix sizes. Once the input matrices are formed using the collected data, the TensorFlow Lite model is run on the data, forming a binary classification that can be presented back to the user.

The overall process therefore forms part of a larger loop as part of the work, where this process loops until the recording is ended by the user. Should the process therefore need to be modified or analysed for its effectiveness, the timestamps of each recording should be able to confirm both the capture time and if any latency is caused by the edge processing.

6.4.3 Smart Home Device

The smart home device was implemented on a Raspberry Pi 4, to simulate a smart home device. Raspberry Pi was chosen due to similar specifications, performance, and size to mid-range smart home devices. The implementation of the code on the Raspberry Pi was completed using Python and Raspberry Pi OS. The Raspberry Pi enabled existing tools, technologies, and peripherals to be used as part of the project. Although the simulated smart home device using a Raspberry Pi is not a perfect solution, existing smart home devices are often closed, meaning that these would not be possible as a test method. The remainder of this section identifies the specifications of the system, the setup of the system, the development process, and the features of the system.

6.4.3.1 Specifications

The smart home device was developed using a Raspberry Pi, which is an inexpensive computer built on the Linux operating system by the Raspberry Pi Foundation. It was developed with the aim of providing a portable, cost-effective, and adaptable platform for teaching computer science and programming to students and enthusiasts. The Raspberry Pi 4 is a single-board computer developed by the Raspberry Pi Foundation. It was released in June 2019 and offers a number of improvements over the previous version, the Raspberry Pi 3 Model B+. The

technical specifications of the Raspberry Pi 4 models are presented as follows:

- CPU: Broadcom BCM2711 quad-core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz
- GPU: Broadcom VideoCore VI
- Memory: 2GB, 4GB, or 8GB LPDDR4-3200 SDRAM (depending on model)
- Storage: MicroSD card slot (up to 2TB)
- Audio: 3.5mm audio jack, HDMI audio
- Video: 2x micro HDMI ports (up to 4Kp60)
- Networking: Gigabit Ethernet, 2.4/5GHz 802.11b/g/n/ac WiFi, Bluetooth 5.0
- USB: 2x USB 2.0 ports, 2x USB 3.0 ports
- GPIO: 40-pin GPIO header
- Power: 5V/3A USB-C power input

The requirements identified during the initial research as part of the development required a device with an ARM processor, which meant that the Raspberry Pi 4 was the selection for the project. In addition to the Raspberry Pi, some additional technologies were used, including an omnidirectional microphone, which is a type of microphone that is designed to pick up sound equally from all directions. Omnidirectional microphones are characterised by their spherical shape and by having a uniform frequency response in all directions. The nature of the algorithm means that to form the most accurate recordings, an omnidirectional microphone should be used when possible, especially within a domestic environment. When compared to other microphone types, the omnidirectional microphone performs better in larger environments or locations where multiple people will be speaking.

Finally, a touch screen and a case for the Raspberry Pi were used. The screen allowed visual feedback to be presented to the end user and interacted with using either on-screen prompts or the included stylus. The screen was a 320x480 pixel

LCD display, which could display the output of the code, or a simulated logo, representing the smart home system that the project was looking to simulate. The screen also contributed to the overall project, providing visual feedback during the design and development stages, indicating how the algorithm was running and determining if any errors were occurring in the processing or detection of violent language.

6.4.3.2 Programming Environment

To develop the smart home device on the Raspberry Pi, Python programming was used on the Raspberry Pi OS. Python was selected because of the wide array of supporting packages available, and the initial TensorFlow model was developed using Python technologies. Python also supported the use of TensorFlow Lite on edge Raspberry Pi devices and was effective at completing the remaining required tasks. Due to this, the numpy library was used for basic matrices calculations and ensuring that the data were formatted correctly for the TensorFlow model. The tflite_runtime.interpreter library was used to interpret and run the model, which then ensured that the previously identified .tflite model file could be used on smart home devices.

The audio was recorded using the PyAudio library, which provides a method of recording audio segments at 10 seconds in length, aligning with the initial dataset. Therefore, the PyAudio library enabled audio recordings to be made, stored, and accessed within the smart home device through the Python interface. Transcriptions were produced using the SpeechRecognition library, which enables multiple different transcription methods to be used; this includes Google Speech Recognition, Google Cloud Speech API, Microsoft Azure Speech, in addition to a number of offline-capable solutions. The SpeechRecognition library therefore enabled the transcriptions to be performed both offline (for edge processing) and online (for higher performance), as required. The nature of the user-based device meant that only a limited number of libraries were required to work with the tflite model, which meant that the overall size of the developed system was minimal when not considering the storage of audio files.

6.4.3.3 Development

To begin the development process, the initial model was transferred to the Raspberry Pi, enabling the previously defined main software loop to be implemented on the system. The development of the system focused on implementation of the overall software loop, with the audio recordings being captured and saved using PyAudio, and the pre-processing stage of transcribing the data with the SpeechRecognition library was completed. This processing was then transformed into the correct sized arrays, which enabled the lite model interpreter to run on the data. The initial stage had little complexity, due to mainly focussing on the audio and data transformations, which have already occurred within Python. Figure 6.2 presents the complete output of this system with two different types of input, showing the outputs of the overall method on the edge device, with the presented transcriptions and the audio being saved.

Figure 6.2 also presents how the initial completion of the model setup was then further integrated into a main software loop. This was initially completed by implementing a timed loop, which was instructed to run concurrently at 10-second intervals; this ensured that while data was processing in the background, the loop would still be able to queue and capture the recordings as required. The initial stage of the loop was the audio recording, from which a saved audio segment from the omnidirectional microphone would be stored based on the timestamp of the recording. This audio segment would be processed as a transcript, from which BERT vectors and LIWC rankings would be formed. The audio aspect of the data would be processed using feature extraction and input into the relevant input shapes for the time domain and frequency features. The local version of Tensorflow Lite was then used to run the model and the BERT, LIWC, time-frequency Domain, and MFCC features were fed into the input shapes. The model was then used to perform binary classification on the verbal and conversation utterances. At present, the classification and edge-based normalisation is presented to the user in text form, and can be stored for future analysis by a trusted individual or trained support officer. The formula for calculating the normalisation of the array can be calculated as follows:

$$(x) = \frac{x - \min}{\max - \min} \tag{6.1}$$

Where x is the input value of the matrix, min is the minimum value, and max is the maximum value of the array.

Finally, advanced functionality was included; this includes adding functionality for output of basic text to the portable screen and the presentation of the values and transcripts back to the user. In addition, the model and related code was set to run immediately on the device starting, which means that the user could then organise the recording as and when necessary, as opposed to having to specifically turn on/off the device when needed. The development was then concluded by ensuring that the BERT vectors were the same as those processed during the initial development of the BERT model, ensuring that the input and output shapes matched the word embeddings.

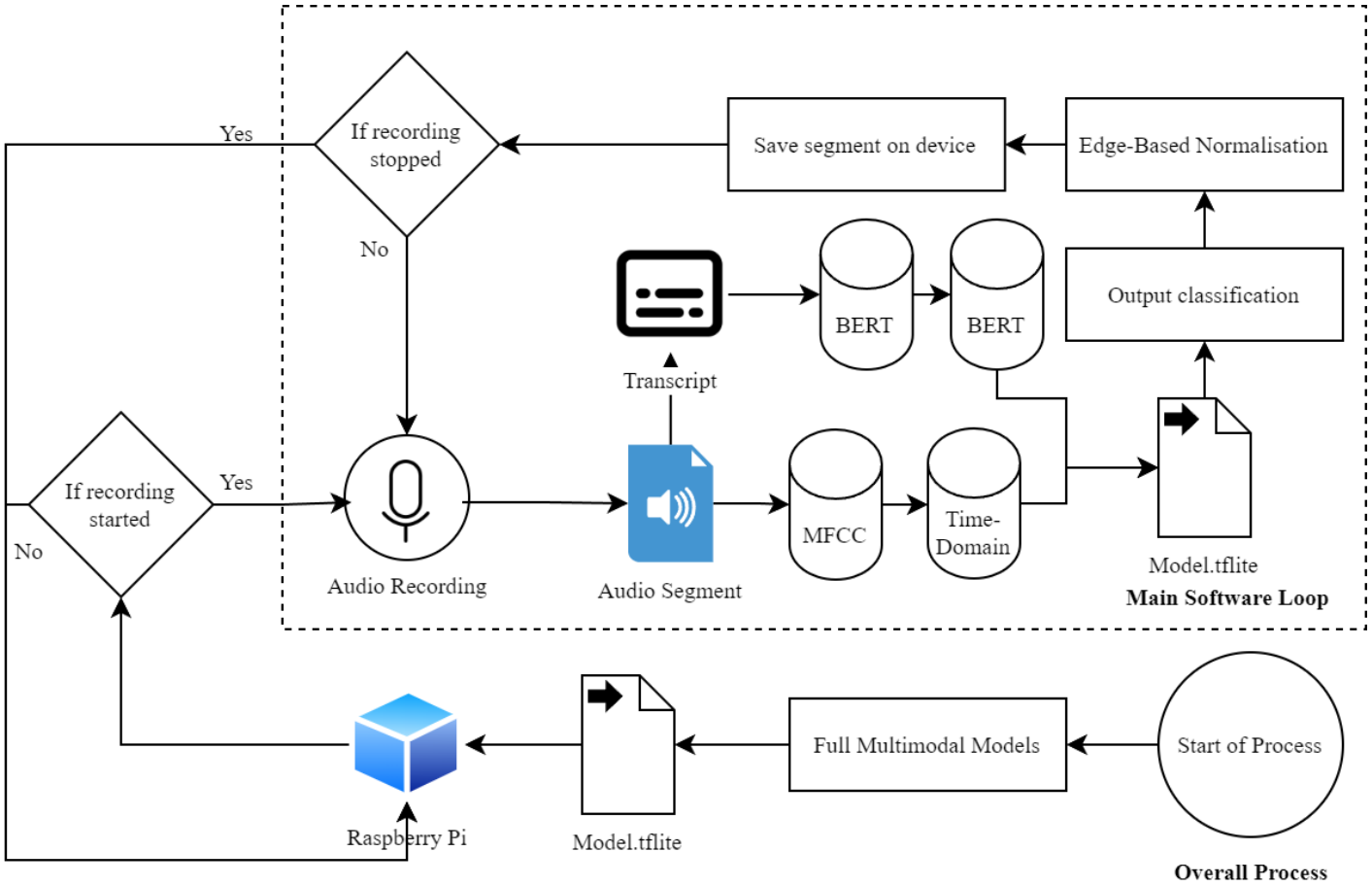

Figure 6.2: An overview of the full fusion model applied to the Raspberry Pi edge device. The diagram initially presents the overall process, including the transfer to a lite model, before highlighting how the main software loop is used within the smart home environment.

6.4.4 Mobile Device

The mobile application was designed to accompany the previously described TensorFlow Lite model. The mobile platform provided a novel challenge due to limited processing power and access compared to other formats of devices. Despite this, the mobile platform provided the best method to ensure ubiquitous access to devices and the highest probability that at-risk individuals use the technology for which they already have access to. The mobile device was in contrast to the smart home device due to the location and use of the device; the mobile device would usually be used by a user on the go with a speaker-facing microphone, or be on their person.

6.4.4.1 Specifications

Using Flutter as a development tool means that iOS, Android, and web-based devices would be supported by the system. Flutter is the deciding factor as to if a mobile platform will be supported by the developed prototype, Flutter supports Android SDK versions between 21-30, iOS version 16, and web-based technologies on modern web browsers. The platform was tested on the Pixel 3a x86 mobile emulator. For practical use, the device was also tested on a Samsung Galaxy S10 + for evidence of the working platform on a physical device. However, the platform specifications did mean that no guarantee of microphone quality could be ensured, meaning that this would largely be based on the hardware of the device, which would need to be verified before widespread use.

6.4.4.2 Programming Environment

The development of the mobile application was completed using Flutter, with Dart used as the main programming language. Flutter enabled cross-platform development to be performed easily when compared to Java and Swift, additionally the TensorFlow Lite Flutter plugin enabled custom TensorFlow models with multiple inputs to be ran on the edge. Despite this ease of use, Dart did make array manipulation difficult to perform, which was required due to the complex input shapes used by the multimodal algorithm, with multidimensional arrays being used for most data storage, specifically in the context of the audio modalities.

However, most of the data for the input in these modalities was easily accessible through the Flutter APIs, which provided access to the on-device audio, file storage, and notification systems.

The complex nature of the application required several packages to be used during the development process. The main package used was tflite_flutter [266], which enabled the interaction with TensorFlow Lite models using Dart. The application also used the record [267] package, which provided access to recording audio in 10-second segments. The application used the Flutter eventify [268] library, which allowed smoother communication of the outputs to be displayed on the application user interface. The matrix2d package [269] was used to perform matrix operations. The libraries highlighted in this section were used alongside other supporting technologies, which allowed the mobile application to be quickly prototyped throughout the development process.

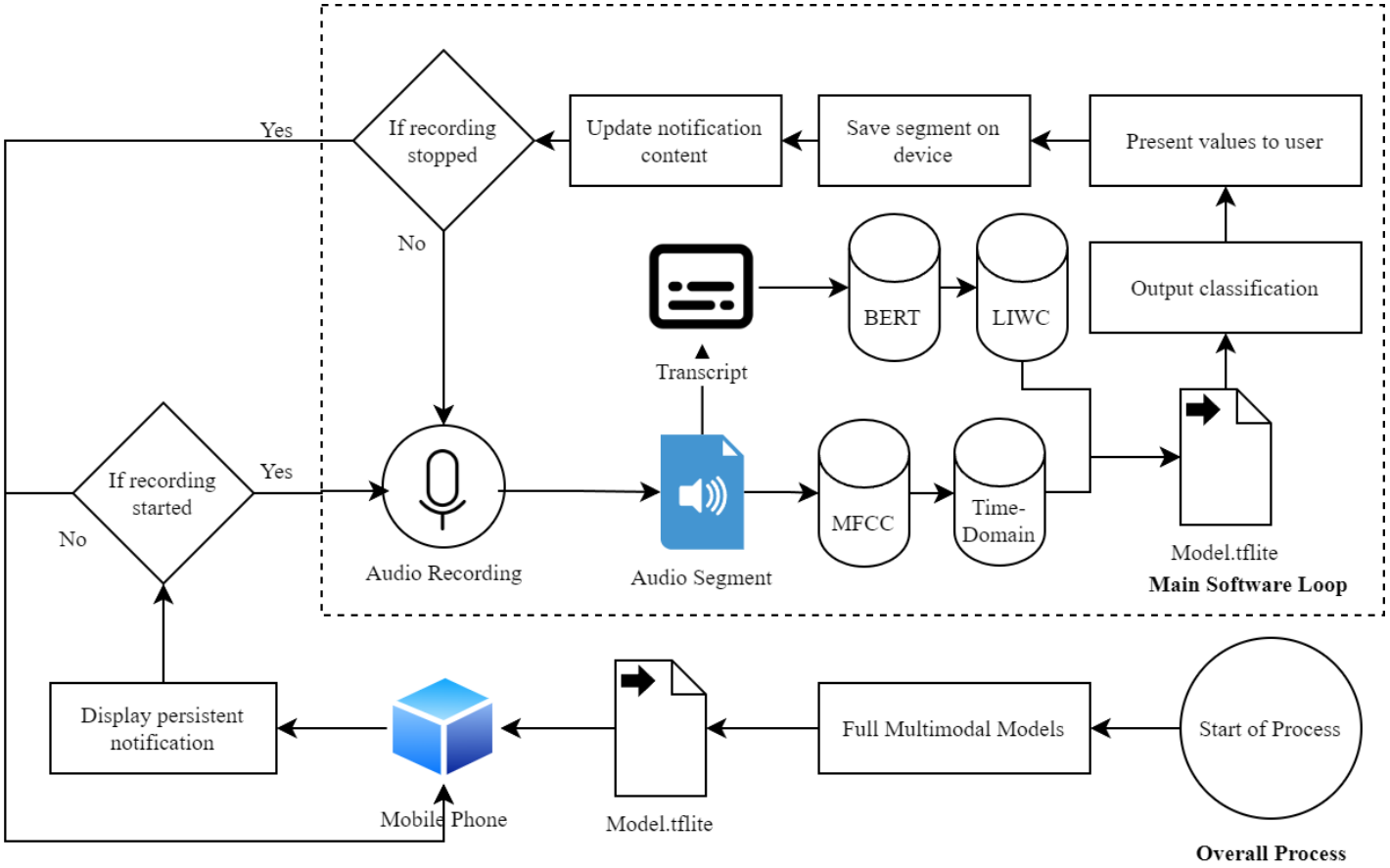

Figure 6.3: An overview of the full fusion model applied to the mobile phone device. The diagram initially presents the overall process, including the transfer to a lite model, before highlighting how the main software loop is used within the mobile environment.

6.4.4.3 Development

The mobile application was developed as a standalone system compared to the smart home system, which could use the same code as the initial Python processing. Figure 6.3 presents the overall processing flow of the mobile application, which is first started when the user opens the mobile application. An overall processing loop was implemented in Dart to ensure that the audio recordings and processing can occur in a multi-threaded process. The main operation loop was activated through two different methods: interaction with the large on-screen button, or through a persistent notification displayed to the user once the app had been opened. The notification selection was attached the users first interaction with the application, which is also the first time Android allows such a notification to be attached, upon interaction with the notification it can either be hidden or selected to start recording. The notification code was asynchronous, so that it could run at the same time as other processing within the application, which is important to consider when working with visual, processing, and notification threads.

Five files were used for the development of the Flutter mobile application, these were organised as follows: (1) the main file contained the main application code such as launching the application and importing the other required files, (2) the timer loop file managed the 10-second timer, (3) the sound recorder which manages the recordings and end of sounds in addition to storing the output file, (4) the classification file which processes the matrices and interpreter result, and (5) the notification manager which is required for notifications to be displayed. Therefore, the overall application is much larger than the one implemented for the smart home device, mostly due to the increased set of features and the visual elements of the application.

Following the main software loop, the application used a 10-second loop implemented in Dart, the loop can be activated using the notification service, or by selecting the button on the main screen. This button, the application theme, and the text content change based on the state and the result of the recording. This was then linked to the sound recorder file, which would activate or end the sound recording and save this to local storage. The end of the recording would

be saved along with a timestamp and stored on the user's mobile device. User data such as the sound recordings is tricky to store; thus, user application data storage is chosen as the location for recordings to be stored, often hidden from end users, while also accessible to be deleted or removed as needed.

To perform the classification task, the file is provided with data in the form of a timestamp which can be used to identify the location of the file. Using the sound file, the matrices storing the data for the model are formed, and the local speech to text processing can be attempted, an online service can also be used due to issues with speech-to-text when using Flutter. The transcription is then also transformed into the required matrix format and stored in an array for processing. Upon the completion of these tasks, and assuming no queue in processing, the matrices and features are interpreted using the TensorFlow Lite interpreter. The interpreter performs the processing on the edge and returns the values required for binary classification; using these values it is then possible to understand the model prediction of if a violent conversation has occurred or not. The results of the classification are then emitted from the class into the main code, updating the visual elements of the application such as the background colour and text, and providing other visual feedback to the user.

To enable feedback and recording while the application is closed, a persistent notification is used, which can provide feedback from the application without the need to have the application directly open. Using the notification service, it is possible to use the eventify package to inform the user of recent results, or enable an emergency recording to be started. The application loop in either case continues to run and enables 10-second recordings to be captured, processed, and stored as part of the application. Design considerations at this stage include the storage of data on the local device, the local edge processing of the complex algorithm, and the potential for close contact or police warnings to be sent during this process.

The overall loop was preserved and the classification would effectively be performed. Some issues were identified during development, which required careful consideration when implementing. For example, the interpreter would sometimes have a small delay while processing the recording due to latency in the mobile device implementation; to prevent this issue from occurring, the application would

queue the required classifications, ensuring that the order based on timestamps is preserved. Another issue identified during the development process was the varied support for onboard speech-to-text transcriptions to be performed on some devices, due to this, and the common connectivity of mobile phones, a backup service of using a speech-to-text cloud service is provided, to ensure an accurate transcription is formed and a more accurate violence classification is provided.

6.5 Results

To test the research artefacts developed as part of this project, it was necessary to test both the smart home device and the mobile device and provide proof-of-concept versions of the system. Unfortunately, due to the ethical limitations of the study, it was not possible to perform user testing at this stage; instead, each of the applications is presented as a demonstration of the features using text-based descriptions and screenshots of the working application. In addition to this, a brief overview of the testing performed is provided and any notable errors or changes made are explained.

Both systems provided an effective method of conducting violent language detection on user devices, the smart home device was able to effectively record, process, and present the results of the violent language classification, while also allowing the processing of the extra features such as local value normalisation to ensure values are contextual to the system being used and to dynamically change the system values based on the known recorded data. For the mobile device, a complete and useful mobile application is provided, in which the user can interact with the model and its data effectively. The mobile application enables the user to capture violent conversations using either the main control interfaces, or the persistent notification that was displayed. The mobile device also provided a useful method of presenting the data of the system, running on the edge, for the purposes of demonstration through animating and modifying the user interface based on the violence detection rating.

Figure 6.4: The smart home device application running as a Python application on a Raspberry Pi. The image displays the resulting system on the touchscreen local display, the omnidirectional microphone, and an input device for the purposes of testing.

6.5.1 Smart Home Device

The resulting application for the smart home device is run-able on a Raspberry Pi, due to the nature of TensorFlow Lite, it is possible to run the application on any platform that supports the package. The produced application is provided as a standalone system, with minimal required packages to be imported as part of the software. The nature of the smart home device system means that minimal visual representations are provided as part of the system; however, Figure 6.4 presents the resulting system, highlighting the interactive touch screen display, the omnidirectional microphone, and the Python code running within the application. The system overall is small and relatively cost-effective to set up, with the only required technologies being the Raspberry Pi and the omnidirectional microphone, which are both widely available and connectable through USB.

The resulting system is a useful method of engaging with the model on edge devices, and while, as shown in Figure 6.4, the device is not visually engaging, the system is built to match a smart home device, which often does not have a display, works in the background and only allows engagement through a mobile application. Therefore, the result of the smart home device is that it is a

good method of running the model on the distributed devices, with performance benefits due to the powerful system being used. The downside of this method is that the system produces limited visual feedback and can be difficult to control, but assuming that the system would be used to captured violent language and conversations and only be run when needed in the background, this would be an effective method of capturing such data.

6.5.2 Mobile Device

The developed application provides an effective method of running the novel multimodal algorithm on an edge device, the mobile device method was implemented using Flutter and Dart and therefore these technologies allow the application to run across Android, web, and iOS devices. The application can also be applied to tablets and other smart devices that are compatible with these technologies, or that support Flutter as an implementation method. The mobile application is the most effective method produced during this project, due to the well-performing algorithm on edge devices and the contextual user feedback presented while the application is running. For this purpose, the mobile application is very useful, as it enables the overall system to present visual feedback while in operation.

The basic application flow, using the same method as for the smart home device, follows the planned software loop. For the mobile device, the activation is caused by a notification or the selection of the main operation button, as presented on the opening screen of the application in Figure 6.5. The main button is animated, and is also the main feature of each screen in the application, with the other content maneuvering around the button when being displayed, to ensure that the user is able to activate the recording, or cancel the recording at any stage during the process. Once the recording has started, the user is shown a small toast notification, and the icon changes to present that the system is now recording; in addition to this, a small animation is played to visualise that the system is currently recording in the background. This animation is featured in Figure 6.6, and is also evidenced in the change icon between the two figures.

After the first 10-second recording has been completed, the visual feedback on the application changes. Background and text-based description changes ac-

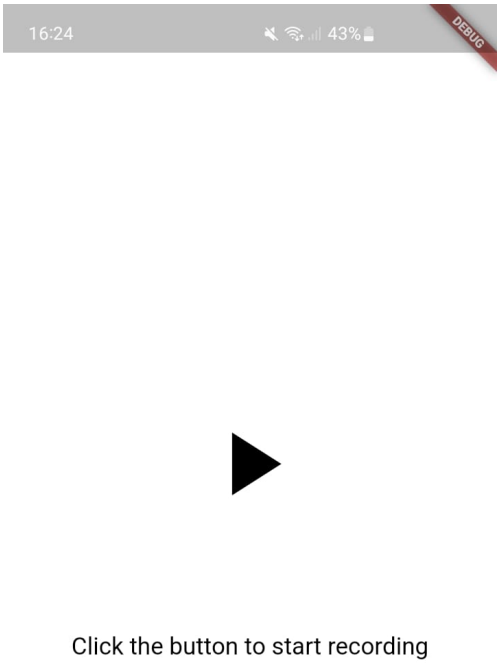

Figure 6.5: Screenshot of the mobile application for detecting violent language. The screenshot displays the initial application opening screen that contains a button which can be used to start the recording.

Figure 6.6: Screenshot of the application, displaying the main screen upon the button being tapped. A visual change in the icon and animation is presented to the user which occurs while the application is recording.

cording to recorded content and the produced classification. This enables rapid understanding of the recorded results and also presents the determined violent language rating to the user. Figure 6.7 presents how this is presented when the recording is classified as potentially being violent language, similarly, Figure 6.8

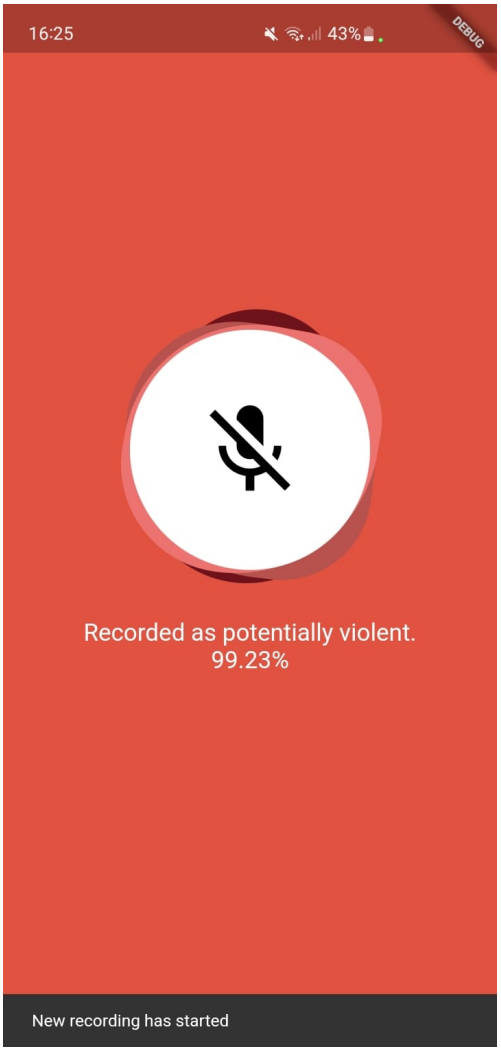

Figure 6.7: Screenshot of the application displaying a red background after a recording loop has been detected as potentially violent. A percentage and text description of the potentially violent result is displayed.

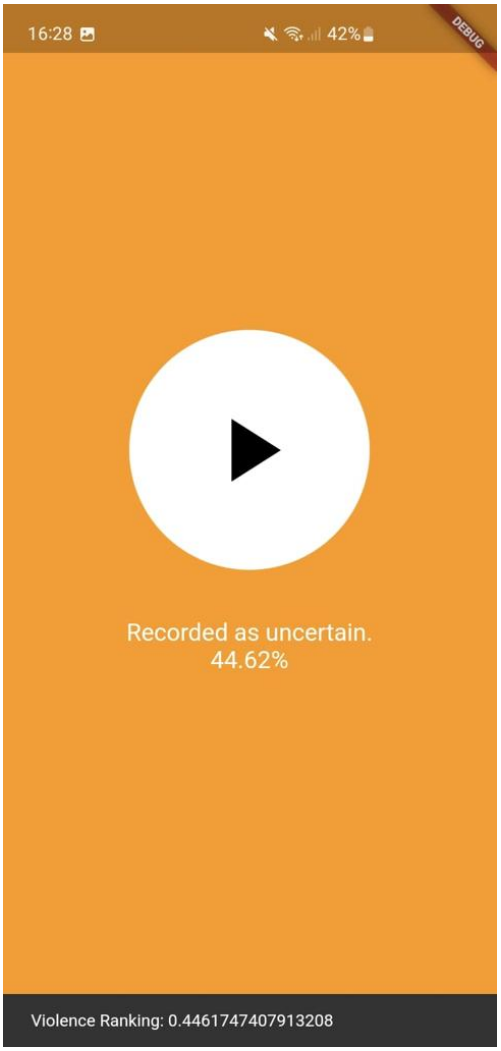

Figure 6.8: Screenshot of the application displaying an orange background after a recording loop has been detected as uncertain. A percentage and text description of the uncertain result is displayed.

presents an orange background and related text when the recording is determined to be within the midpoint of the classification, and finally, Figure 6.9 presents the user interface when the model is determined to be normal conversation, with no violent conversation captured. Although the classifications are not perfect, based

on the previous results these are fairly accurate, especially when considered as a group of values.

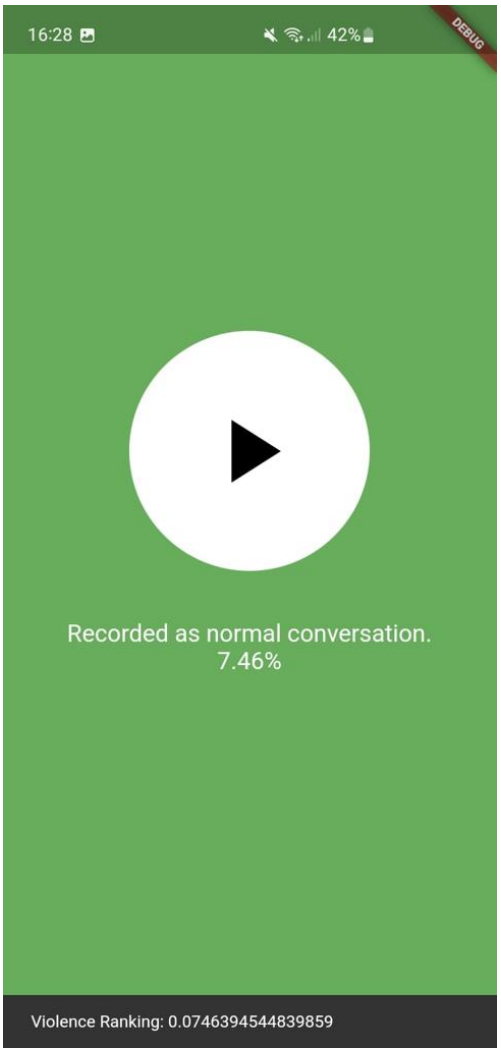

Figure 6.9: Screenshot of the application displaying a green background after a recording loop has been detected as normal conversation. A percentage and text description of the normal conversation result is displayed.

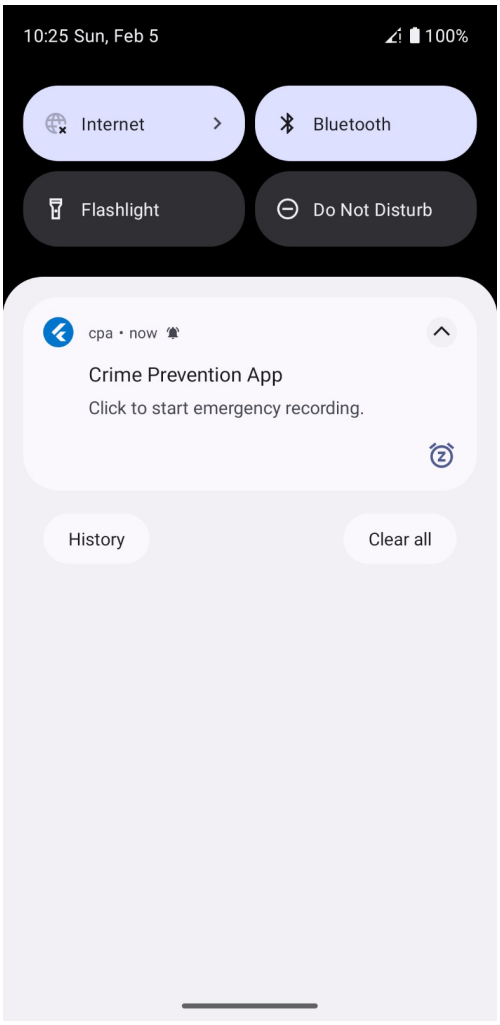

Figure 6.10: Screenshot of the user notification menu displaying the persistent notification from the application. When this notification is selected a recording loop is toggled with the application.

Figure 6.10 presents the notification display, which is displayed to the user once the application has been activated in the background. Selecting the notifica-

tion allows the user to immediately start recording, should a violent conversation start occurring or the user feels like they are in danger. The notification is an effective method of matching some of the design considerations, due to the importance of control in the application and the importance of the application having some form of notification that is persistent while running.

Overall, the proposed system was effective in classifying violent conversations on the edge, with the same model used as the full TensorFlow implementation. The latency on the mobile application is minimal; however, a queue-based system was implemented to support the classifications being performed in the correct order. The overall system also provided good feedback to the user, which is covered by changing the background, identifying the recording, and providing the notification and button control to the user. The mobile application could be further improved by adding additional features, such as location or other modalities, to the dataset, but this would require changes to the overall model.

6.6 Discussion

The application has successfully demonstrated that a novel multimodal algorithm can be processed in near-real time on an edge device. Furthermore, the contextual application area for the detection of violent conversations has also been shown to work effectively on a mobile device. The area of mobile devices presents an interesting research scenario, due to the ubiquitous nature of the devices and the accessibility to a user in domestic environments. These domestic environments are expected to be the main area of use for the application; however, there is potential for location-based data to be used in the future to attach extended contextual information to the application (e.g., geofencing, location alerts, Bluetooth sensing) [270]. The application presented positive aspects of running the algorithm on the edge, with the potential for recordings to be more accurate in the future by curating a larger dataset of audio recordings that occur on a similar microphone to one that could be found on a mobile device.

Several potential issues were discovered during the design and development stage of the application. First, the processing power of the mobile device should be considered due to the time required for both audio and text modalities to

be detected and converted. It will still be possible for the model to run on less powerful phones; however, the queue system used to manage the order of processing may become backlogged. Potential users and security services should also consider the accuracy of the overall algorithm, despite the high accuracy rate of the contextual situation reported [271]. However, this high accuracy is not perfect, which means that extra precautions should be taken to ensure user safety while using the application. This also highlights a need for further development using the multimodal model to occur, encouraging future work to focus on gaining the best possible accuracy alongside the application design giving the user overall control and reporting abilities.

Considerations will need to be made should the application be further developed. In addition to improved accuracy and features, should the application be used by Police to collect or capture data, improvements will need to be made with respect to data storage. While existing data is encrypted in the users' local storage to ensure privacy, should the tool be used as a method of capturing evidence, changes to the security would need to be made for this. For example, a secure key may need to be generated for access, and a log of deleted recordings and times may need to be kept to ensure that the data collected can be used as evidence in a policing situation.

Both applications provided different functionality; the smart home Raspberry Pi was the initial implementation, which initially took a long time for development. The complexity of the applications is also different, the Python implementation has more complexity in code, such as contextual normalisation of the recording values, however, the mobile application is more complex in terms of overall features and the amount of actions happening within the application. The mobile features such as the feedback and the notification do make the mobile application the more impressive in the context of features.

6.6.1 Challenges and Opportunities

The violent conversation crime prevention application was designed to prioritise safety and security. The application was designed to protect user privacy through processing all data on the edge, specifically on the end-users' device. This enabled

all potentially sensitive data to be accessible only through the end-user mobile device, which is often encrypted; however, future investigations may need to consider devices that are not always encrypted. User data is collected through the on-device microphone, data is stored on the local device, and only used for the purposes of detecting violent conversations. Should data need to be transferred in future iterations, it will be necessary to consider encryption and secure transfer of user data, which is a common concern in edge literature [257]. At present, the application does not contain the ability to transfer transcriptions or recordings, and therefore only the resulting binary classification could be transferred.

Further consideration needs to be taken to ensure user safety while using the application; this might include an option to hide the application under the name of a different application such as a chat or messaging device. Therefore, this option would allow the application to share the name, icon, and interface of a different application, with recordings occurring then without the victim opening the named application.

Future work should look to gain ethical approval for investigating the use of such systems with users, although this is out of scope for the research conducted as part of this work. Improving the testing stage through user studies would be the next step in the process of achieving results through the edge application. As identified in previous literature [262], mobile crime prevention applications need a stronger evidence basis for design and implementation. The model implemented works well, but to analyse how the users interact with the functionality would be interesting, in addition to exploring how the privacy of the data implemented in a edge system could support user confidence in this regard. Therefore, further work should look at taking the application to workshops with experts or user studies to investigate this as the next step in the process.

6.7 Conclusion

The application presented in this article was an effective method of running a novel multimodal audio-text model on the edge. The work highlights the potential of distributed and user processing for sensitive topics such as domestic violence and crime prevention. The work considered the design considerations

needed for working with sensitive data, while also developing a solution to run a complex machine learning model. The application could be improved by increasing accuracy, which could be through more realistic recordings on devices with similar microphones.

Future work will now focus on fine-tuning the algorithm and the application design considerations. This could include lab-based testing scenarios, with the hope of future work that would then test the application in domestic environments. Due to the sensitive nature and subject area of the application, these domestic environments can be simulated with expert interviews to determine the effectiveness of the application. Overall, the work will aim to measure the application in both qualitative and quantitative scenarios, enabling effective results to be reported.

Chapter 7

Conclusions and Future Work

7.1 Chapter Overview

This chapter serves as the culmination of the research project on the use of technology in crime prevention, providing an overview of how the objectives were met and carried out successfully. The chapter delves into the objectives of the work, including dataset curation, NLP, multimodal violent language processing, edge processing of violent language, legal and ethical considerations, as well as the discussion of findings related to violent language detection and edge processing. As the final chapter, it offers a comprehensive discussion of the research, summarising the different contributions, providing recommendations, identifying potential areas for improvement, and outlining conclusions and directions for future investigation.

7.2 Review of Aims and Objectives

The introduction of this thesis provided an aim and five objectives, which are critically reviewed with respect to the topic. The aim and objectives were followed throughout the project, to ensure that each planned aspect of the study was completed on time and effectively. The aim was defined and reviewed as follows:

1. **To analyse the data fusion of multimodal sources to determine if a conversation contains violent or aggressive content on the edge.**

The aim of the research, which was to analyse the use of data fusion of text and audio sources to determine if the conversation was violent or not, was successfully achieved, the multimodal and single model attempts were all effective at analysing the content and context of conversation that formed part of the dataset. The evaluation of results and the discussion of each section, each highlighted how the models implemented were effective, and compared the techniques to baseline measures and previous work. The application of the model at the edge was also effective, as the model was able to record and process the audio and text features effectively using edge technologies, including mobile applications and smart home devices. Overall, the aim of the project can be considered to have been met during the course of the research.

The objectives form the more specific aspects of the study, which were defined and reviewed as follows:

1. **To perform a review and evaluation of the existing literature, sources, and applications relating to the use of technology and Ai in crime prevention.**

   The objective of conducting a review and evaluation of existing literature, sources, and applications related to the use of technology in crime prevention has been achieved successfully. A comprehensive review of scholarly articles and publications in the field was conducted that included various technological approaches used in crime prevention, this includes CCTV, data analytics, biometrics, AI, and IoT devices. The evaluation process involved a thorough analysis of the strengths, limitations, emerging trends, challenges, and ethical considerations associated with these technologies. Through the synthesising of the findings, valuable insights were gained, gaps were identified, and areas for further research were proposed. This comprehensive review provides a deeper understanding of the effectiveness and potential impact of technology in preventing crime.

2. **To curate a multimodal dataset of linked audio and text features that can be employed to test and train data fusion models.**

The objective of curating a multimodal dataset of linked audio and text features was successfully achieved for the testing and training of data fusion models. A comprehensive dataset was collected, organised, and prepared consisting of paired audio recordings and corresponding transcriptions or textual representations. The audio data was sourced from British TV series, capturing conversations or speeches that contained screaming and abuse, while the text data were meticulously aligned with the content of the audio recordings. Through careful curation, the dataset ensured accurate linking and annotation of the audio and text pairs, allowing effective training and evaluation of the data fusion models. This curated dataset served as a valuable resource, allowing the evaluation of data fusion models to integrate audio and text information for tasks such as detecting violent or abusive language and sentiment analysis.

3. **To investigate the results of natural language processing feature extraction, and determine the most effective combination of text-processing models.**

   Investigating combinations of NLP approaches to identify the most accurate method was successful in achieving its goal. Performance metrics were used to compare the effectiveness of different models such as LIWC, BERT, and GloVe through methodical experimentation and careful analysis compared to the baseline models and existing work. Through evaluation, the best combination was found to accurately identify violent or aggressive language using BERT and CNN, while LIWC was deemed useful for fusion approaches. This discovery directs the development and use of the model during the later phases of the research.

4. **To develop a data fusion model that combines audio and text modalities to accurately identify violent language from conversations.**

   In the research project, a robust data fusion model was built and improved using machine learning techniques. To accurately categorise conversations as violent/aggressive or non-violent/non-aggressive, this model combines

text (LIWC and BERT) and audio (MFCC, time-frequency domain) modalities. The model was meticulously trained on a large dataset with labelled annotations, allowing it to recognise patterns and extract pertinent features that are indicative of violent or aggressive language. The overall F1 score of 0.85, among other exact performance evaluations, confirmed the model's high accuracy in identifying this language. Especially in areas such as prevention of domestic violence, public safety, and online content moderation, this achievement offers a useful automated tool for crime prevention.

5. **To perform model reduction on the data fusion model to embed the model on edge devices in real-world scenarios.**

   Through careful optimisation and refinement, the data fusion model was adapted to meet the computational constraints of smart home and mobile edge devices. Techniques such as model compression were used to reduce the size and computational complexity of the model while preserving its accuracy and effectiveness in detecting violent or aggressive language. By developing a lite version of the data fusion model, the research has paved the way for its seamless implementation on resource-constrained devices. This implementation enables the processing and detection of violent or aggressive language in real-time, empowering users to prevent and address potential violent conversations in domestic contexts.

7.3 Discussion of Findings

This thesis presented the overall results of a text-based NLP model, an audio feature extraction model, a multimodal model, and an edge processing device to classify and report violent language or violent conversations. This section will discuss the overall findings of the work, including a discussion on the curation of the custom text-audio multimodal dataset that was produced, the NLP method used and the results achieved, the multimodal and audio processing techniques that were used to further augment the text processing, and the edge processing applications which were produced as part of the work.

This thesis presents a complete overview of the algorithm results reported

during the research project, and a selection of existing models to compare with the work. As presented in the table, the work produced and the final F1 score was the combination of MFCC, time domain, LIWC and BERT characteristics, which achieved a F1 score of 0.85. Compared to existing work, the model is close to the augmented model attempted previously [248], while showing improvement compared to existing fusion attempts, which scored 0.77 [226] and 0.67 [248]. The fusion model also presents an improvement over the initial baseline model, which achieved an F1 score of 0.75 when considering a random forest on all features. The combined model also showcased improvement over all the individual and small fusion models that were produced, with the BERT and CNN model achieving 0.67, the MFCC and time domain feature model achieving 0.80, and the LIWC model achieving 0.65, indicating how the work produced has been effective in achieving this purpose.

7.3.1 Dataset Curation

Finding a suitable dataset for this study was challenging, as there are several issues in the dataset acquisition process that focus specifically on domestic abuse may be a challenge for several reasons. First, domestic abuse is a highly sensitive and private issue that raises ethical and legal considerations regarding the collection, storage, and sharing of data. Second, domestic abuse is often unreported, making it difficult to collect comprehensive and representative datasets. Third, the protection of the privacy and confidentiality of the individuals involved, including victims and perpetrators, is of the utmost importance and leads to the limited availability of publicly accessible datasets. Finally, it is also difficult to obtain informed consent from individuals affected by domestic abuse due to power dynamics and security concerns. Researchers and organisations often establish specialised collaborations or initiatives to collect data in accordance with ethical standards and the priority for the well-being of survivors and victims. However, in this case, it was not possible to obtain actual data and thus it was necessary to create one using scenes from British TV series. As discussed above, the dataset was very well conceived and was labelled and overlooked by professionals.

However, the limitations of the dataset is the size, with a limited number of

classifications included. To further enhance the accuracy of the detection system, the research recognised the importance of a larger training dataset. By expanding the dataset, the models can capture a wider range of linguistic variations and contextual nuances, ultimately enhancing the ability to accurately recognise and classify instances of anger and abuse. This improvement to the dataset could be completed by adding more contextual data to the initial recordings or by implementing the produced edge application as a feedback loop, meaning that the data collected would then be used to further improve the standards and number of recordings available. It would be necessary for future work to further add to the dataset and explore possible methods of capturing more realistic data using a range of smart home and mobile phone microphones. Further work in this regard could therefore support more effective models and encourage more detailed comparisons of results.

7.3.2 Natural Language Processing

The research findings demonstrate the application of NLP techniques in detecting violent conversations by analysing language cues. By incorporating multiple modalities and using two specific feature sets, significant improvements in the accuracy of the detection process were achieved based on the combination of the text features.

The combination of LIWC and BERT played a crucial role in enhancing the results. Where initial F1 results were reported to be 0.67 for BERT and 0.65 for LIWC. The overall model showed a .05 improvement when further combined with the text modalities, highlighting the importance of the NLP on the overall success of the multimodal attempt. The integration of this set of characteristics provided a comprehensive understanding of the linguistic patterns associated with anger and abuse, leading to a more precise identification and increasing the range of characteristics identified about the language through the implementation of BERT and LIWC. Compared to existing text models presented in Table 5.5, the work produced performed well compared to the context-free text model [226] and the weighted text model [226] F1 scores. While the results of the overall work were similar to the sequence text model [226] and the complete transcription and

audio model [248].

The limitations of the natural language processing relate to the size of the BERT vectors and to the performance of the individual LIWC modality. For this to be improved dimensionality reduction could be integrated with the BERT vectors, and additional contextual attributes from LIWC or similar analysis software could be used. The successful integration of NLP techniques and the synergistic effect of combining LIWC and BERT feature sets underscore the contribution of the research. Furthermore, the findings highlight the potential for continuous improvement through the expansion of training data. Additionally, the expansion of the dataset could also contribute to improved outcomes in detecting and addressing anger and abuse using NLP.

7.3.3 Multimodal Violent Language Processing

The results of this study provide insight into the efficacy of multimodal techniques in the identification of violent language to prevent crime. The combination of audio and text modalities with MFCC, time-frequency domain features, LIWC, and BERT produced encouraging results and showed the potential of the research. The full fusion model presented in Table 5.5 achieved a remarkable F1 score of 0.85, which is notable for its improvement over individual modality methods and scored higher than other methods such as the 0.67 [248] transcription and audio method reported previously. The model also showed improvement over the previous literature, through the increase in the score of the 0.77 F1 score reported using both audio and text features [226]. Impressively, the model scored similar to specific models, such as an augmented transcript and audio model, which had an F1 score of 0.87 [248]. This demonstrates the value of combining different modalities to improve the precision of identifying violent conversations. Although they have slightly lower F1 scores than audio fusion and individual audio features, text-based fusion techniques still need to be improved and explored further. Expanding the corpus of transcriptions used in the analysis could potentially improve the performance of text-based fusion approaches.

The main limitation of this finding is the application of the model, where further examination on the effectiveness and applicability of multimodal fusion

techniques in context-specific scenarios, such as actual instances involving violent conversations, should be further explored. This ongoing emphasis on enhancing and assessing the model's performance in real-world situations will advance efforts to prevent crime. It is critical to recognise the constraints of this study, as 15% of the segments were incorrectly classified, the F1 score of 0.85 indicates that there is still room for improvement. Future research can take this into account by using a larger dataset and more contextual information, such as taking into account the number of speakers.

The study's findings demonstrate the potential of the work in multimodal violent language processing for preventing crime. While there are obstacles to be overcome, the encouraging findings and ongoing research in this area provide a strong incentive for more research and method improvement.

7.3.4 Edge Processing of Violent Language

The findings demonstrate the potential of edge processing in preventing crime, particularly when it comes to identifying violent conversations. Through implementing a multimodal algorithm that runs in real-time on edge devices, this method not only demonstrates the viability of edge processing in the context of domestic environments but also ensures enhanced privacy and data security by processing the data locally on the user's device. The use of the algorithm on mobile devices produced encouraging results, proving potency in spotting violent language in real-world situations.

Edge processing in crime prevention has a number of benefits when used effectively. First, it gives users more power by enabling them to keep control of their data and avoid any potential privacy issues caused by centralised processing. Second, it takes advantage of the widespread use of mobile devices to offer a simple method to address violent language in domestic settings. The use of edge processing also creates possibilities for incorporating location-based data, such as geofencing, location alerts, and Bluetooth sensing, to improve the application's contextual understanding and offer more comprehensive situational awareness.

The findings of the thesis highlight how edge processing can revolutionise the field of crime prevention. The result of this finding requires further investigation

and optimisation of the algorithm for the purpose of achieving higher accuracy rates while giving users a full range of control and reporting options. To ensure the application's successful deployment and adoption, model quantisation, processing power, algorithmic precision, and data storage security needs to be considered and studied in user-based scenarios. Edge processing can continue to enable more efficient and privacy-preserving crime prevention solutions by integrating technological advances with user-centric design principles, however, the new approaches require further study.

7.4 Legal and Ethical Considerations

7.4.1 Violent Language Detection

The implementation of violent language detection systems for crime prevention raises important legal and ethical considerations. These include privacy rights, data protection, bias and fairness, potential misuse of technology, legal compliance, and ethical implications. It is crucial to respect individuals' privacy rights and comply with relevant privacy laws when collecting and analysing audio and text data. Strong data protection measures, such as encryption and secure storage, must be in place to protect sensitive information. Addressing biases in language models and data sources is essential for fair outcomes, and regular auditing can ensure equitable treatment. Transparent guidelines are necessary to prevent technology misuse and protect against false accusations or rights infringement. Compliance with applicable laws and regulations, such as data privacy and consent requirements, is necessary. Ethical considerations should guide the development and deployment of these systems, promoting responsible use, transparency, and respect for individual rights.

7.4.2 Edge Processing

Legal and ethical considerations are crucial when using edge processing for a crime prevention tool that records conversations. From a legal perspective, adherence to relevant laws and regulations is paramount. Compliance with data privacy

and protection laws and consent requirements must be ensured. Proper consent from all parties involved would be essential, and collected data should be handled according to privacy rights and legal obligations.

Ethical considerations include transparency, informed consent, and fairness. Considerations about which users would need to be fully informed about the monitoring of their conversations and understand the purpose and implications of the technology. The development and deployment prioritised fairness, accountability, and the avoidance of discrimination or bias; however, this is something that requires further work to improve accuracy. Regular audits and tests would need to be conducted to address any biases or ethical concerns.

Data security is of utmost importance to protect against unauthorised access, breaches, and misuse. Robust measures such as encryption, access controls, and secure storage practices should be implemented. Mitigating false positives, where non-abusive conversations are wrongly flagged, is critical to avoid unnecessary harm or accusations. Proper oversight and accountability structures would need to be established, including regular monitoring, auditing, and review processes, to ensure responsible use of technology and prevent potential abuse or misuse. It should be noted that this technology would only be given out by authorities in extreme cases where the victim would ask for it as a form of security to stay safe from the abuser without any direct involvement from the police.

7.5 Recommendations

In an era marked by technological advancements, the integration of crime prevention devices into everyday technologies has emerged as a promising approach to enhancing personal safety and combating domestic abuse. This section delves into three distinct scenarios where crime prevention can allow victims to remain safe: Public safety policies, user safety, and commercial gains. By leveraging text-audio analysis and sophisticated algorithms, these devices are able to actively monitor conversations and detect signs of domestic abuse, ultimately providing a safety net for individuals in need.

7.5.1 Policy

If used in policy, this research offers a number of compelling benefits that will improve the safety and well-being of victims. The ability to help early intervention is one of its main advantages. Law enforcement can respond quickly to these upsetting situations by quickly seeing indicators of domestic violence in real time, thus protecting victims from additional injury. In addition, the device can improve victims' protection as it guarantees that victims receive prompt assistance and the necessary support by quickly alerting authorities. The ability of the tool to record audio as a method of proof of domestic abuse is another important benefit. These recordings can dramatically improve the chances of a successful prosecution, strengthen the case against offenders, and aid police investigations. The tool helps in the pursuit of crime prevention by offering additional documentation beyond the victim's statement.

The device could also provide important safety safeguards for responding police officers. Before they arrive at the site, real-time notifications provide officers with crucial information about ongoing abuse, allowing them to take the necessary precautions and provide more effective help. However, it is crucial to set precise legal and ethical rules to protect the rights and privacy of everyone concerned. Consent, data protection, and transparency issues must be addressed through policy. For the device to be used effectively while providing the highest level of protection for the interests of the victims, law enforcement professionals must undergo regular training on its responsible usage and ethical considerations.

7.5.2 User

The implementation of a crime prevention device that can detect domestic abuse in real-time and promptly alert the authorities offers significant benefits in improving the safety and protection of victims. By enabling law enforcement to react quickly to domestic abuse incidents, this research offers the possibility of preventing more crime against victims. The research also provides a covert way for victims to obtain help without disclosing their situation to the abuser, acting as a silent distress signal that gives them a sense of security.

The ability of this device to record abusive conversations through audio is

one of the benefits, as it provides the victim with documentation to support their allegations and improves their legal case against the abuser. This supporting evidence could be vital in informing the appropriate authorities about the abuse and supporting a thorough and effective legal procedure. Additionally, the awareness that discussions are being recorded serves as a deterrent, possibly discouraging potential abusers from acting abusively and making the environment for victims safer.

7.5.3 Commercial

In the competitive landscape of consumer electronics, innovation is the key to attracting and retaining customers. Manufacturers are constantly developing new features and functionalities to address urgent social problems. Domestic violence is one of those issues that needs to be addressed due to it being a widespread problem that affects millions of people around the world. Companies have begun investigating how to incorporate crime prevention technology into mobile device such as iPhones and smart home systems. These technologies give an extra layer of safety through recording and looking for indicators of violent conversations.

7.6 Limitations and Future Work

A limitation of this work pertains to the challenging task of detecting mental manipulation within conversations. Mental manipulation encompasses subtle tactics such as coercion, gaslighting, and emotional manipulation, which are not easily discernible through automated audio analysis or text transcriptions alone. The intricate nature of mental manipulation relies on nuanced cues, tone of voice, nonverbal communication, and contextual understanding. These factors pose difficulties for the proposed system in accurately identifying and flagging instances of mental manipulation, potentially leading to missed detections or false negatives. To overcome this challenge, alternative approaches must be explored. Integrating advanced techniques, such as behavioural analysis, could improve the capabilities of crime prevention systems, allowing them to detect and respond to cases involving mental manipulation more effectively.

Another limitation of this study is related to the curation of a labelled multimodal dataset. The dataset consisted primarily of recordings from television series, which may not accurately represent real-life scenarios in terms of audio quality and conversational dynamics. To obtain more precise sound samples, it would be preferable to record data directly from phone conversations or smart home devices. Furthermore, the nature of the research made it impractical to conduct actual testing or validation in real-world settings. This limitation hinders a comprehensive assessment of the system's performance and practical viability in real-time situations. To address these limitations in future research, it is essential to gather a diverse and comprehensive dataset that closely reflects authentic conversations while ensuring participant privacy and consent. Additionally, conducting extensive field testing and validation would provide stronger evidence of the system's capabilities and enhance its practical utility in detecting and addressing abusive conversations.

While the work conducted in this thesis used the F1 score to measure the harmonic mean, the precision and recall of the algorithm used to classify audio as violent or non-violent also presents potential limitations of this research. Limitations arise from the trade-off between precision and recall. Although high precision indicates a low rate of false positives, ensuring accurate predictions of violent segments, it may result in lower recall, meaning some instances of actual violence may be missed. Conversely, high recall captures a greater proportion of violent instances but increases the risk of false positives, classifying non-violent segments as violent. The model results presented in this thesis attempt to strike a balance between precision and recall using the F1 score, although limitations with a binary classification of subjective human language still presents challenges to the work, future work could include collecting a larger selection of data to expand the classification possibilities.

Another limitation stems from the algorithm's performance in specific conditions. Factors such as audio quality, background noise, accents, and speech variations can influence the algorithm's ability to accurately classify audio segments, leading to potential misclassifications and reduced overall effectiveness. Concerns regarding audio quality can also impact the text-based transcriptions, causing problems for both modalities. To address these limitations, ongoing re-

finement and optimisation of the algorithm is essential. This involves collecting various training data representing real-world conditions and continuously fine-tuning the algorithm to improve both precision and recall rates. Rigorous testing and validation procedures should be implemented to evaluate the algorithm's performance across scenarios, ensuring its reliability and practicality in real-world applications.

This research demonstrates promising potential for future development and expansion of crime prevention research that investigates conversations and detects violence or danger. There are several avenues for further enhancement and augmentation of the research. For example, integrating additional modalities, such as Bluetooth, could enable distance tracking between the abuser and the victim, triggering proximity alerts. The current implementation has been carried out on a computer using a Raspberry Pi and a mobile phone, but the next stage could involve transferring the research to wearable technologies. This transition opens opportunities to incorporate more sensors and gather richer data for more accurate classification.

By deploying the research approach on a wearable device, such as a watch, could assist in getting closer proximity to the microphone, reducing the risk of missing crucial information during conversation analysis. Additionally, a wearable device offers the potential to integrate a heart rate sensor to detect patterns of distress and assess the impact of the surroundings on the individual. These advancements will contribute to a more comprehensive and nuanced classification process, enhancing the overall effectiveness of the crime prevention tool.

7.7 Conclusion

This thesis presents the use of digital technologies to process, identify, and classify violent language on edge devices. The work carried out throughout the investigation has provided a novel methodology for working with complex, large, and multimodal datasets on edge processing devices such as mobile phones and smart home technologies. The overall results of the project can be considered a success, due to the effective demonstrations of individual modalities, the positive results of the multimodal algorithm (F1 = 0.85), and the technical complexity of the

edge processing applications for mobile and smart home technologies. Therefore, the results of the overall project have been positive, with the work conducted effectively investigating the use of edge technologies to detect violent language.

Techniques such as BERT, LIWC, time-frequency domain feature extraction, and MFCC were effective in being used for the purposes of violent language detection. The BERT and CNN model, for example, achieved an F1 score of 0.67, which, although still low, was then further improved upon being combined with the LIWC, MFCC and time-frequency domain features, successfully achieving an F1 score of 0.85, a big increase on the single modality studies that were performed. This improvement in results shows how the techniques used in this thesis were effective in improving the classification of violent language in conversations through the use of multimodal models. The resulting algorithm, while effective, could be further improved by implementing a larger data source, meaning that both the audio and text modalities would have improved results during the fusion stage. However, retaining the relatively low size of the model did mean that the overall version could then be implemented on edge devices, demonstrated in this work as a mobile application and a smart home device. Therefore, the developed implementations could be applied as examples of using contextual but complex multimodal models on edge devices through model reduction.

In general, this technique would be an effective method of classifying violent language to prevent and deter domestic violence; this is due to the proposed application that forms an evidence capture tool which could be used for crime prevention, safety enhancement, and evidence collection. Techniques, methods, and research conducted as part of the Ph.D. have highlighted how such methods could be used in other contextual examples, such as detecting depression, violent language online, or violent actions. Overall, the research conducted was able to successfully apply a text-audio fusion technique using BERT, LIWC, MFCC, and time-frequency domain features to achieve an F1 score of 0.85, and then apply this rich multimodal algorithm on edge devices. Edge devices such as mobile phones and smart home devices were tested, which evidences how the contextual study scenario could work, due to the domestic nature of such devices. The work presented in this thesis can be further expanded into other disciplines and contextual study areas, but future work should focus on three main activities:

(1) curating a larger dataset of violent conversations with both text and audio features, (2) developing more contextual models to explore how the results of this work could be applied to other study areas, and (3) testing edge devices through user studies, to analyse how users interact with such technologies in a HCI approach. Therefore, conducting this future work would allow technology to be further evaluated for its potential use as a crime prevention tool.

References

[1] Apple. Use emergency sos on your iphone - apple support uk. https://support.apple.com/en-gb/HT208076, 2023. xiv, 36, 37, 38, 39

[2] Nor Haizan Mohamed Radzi, Haslina Hashim, et al. Research on emotion classification based on multi-modal fusion. *Baghdad Science Journal*, 21(2 (SI)):0548–0548, 2024. xvi, 122, 123

[3] Zhiqiang Gan, Xiang-e Sun, and Meihua Liu. A multimodal sentiment model based on causal gating attention mechanism. 2024. preprint. xvi, 123, 124

[4] Lynnmarie Sardinha, Mathieu Maheu-Giroux, Heidi Stöckl, Sarah Rachel Meyer, and Claudia García-Moreno. Global, regional, and national prevalence estimates of physical or sexual, or both, intimate partner violence against women in 2018. *The Lancet*, 399(10327):803–813, 2022. 1, 2

[5] World Health Organization. Violence against women. https://www.who.int/news-room/fact-sheets/detail/violence-against-women, 2021. 1, 2

[6] Claudia García-Moreno, Christina Pallitto, Karen Devries, Heidi Stöckl, Charlotte Watts, and Naeema Abrahams. *Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence*. World Health Organization, 2013. 1

[7] Verena Kolbe and Andreas Büttner. Domestic violence against Men-Prevalence and risk factors. *Dtsch Arztebl Int*, 117(31-32):534–541, 2020. 1

[8] Jan van Dijk, Paul Nieuwbeerta, and Jacqueline Joudo Larsen. Global crime patterns: An analysis of survey data from 166 countries around the world, 2006–2019. *Journal of Quantitative Criminology*, 38(4):793–827, 2022. 1

[9] Anant Kumar. Covid-19 and domestic violence: A possible public health crisis. *Journal of Health Management*, 22(2):192–196, 2020. 1, 2

[10] Alex R. Piquero, Wesley G. Jennings, Erin Jemison, Catherine Kaukinen, and Felicia Marie Knaul. Domestic violence during the covid-19 pandemic - evidence from a systematic review and meta-analysis. *Journal of Criminal Justice*, 74:101806, 2021. 1

[11] Meghan Elkin. Domestic abuse in england and wales overview: November 2022. https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/domesticabuseinenglandandwalesoverview/november2022, 2022. 2, 3, 17

[12] Meghan Elkin. Perceptions of personal safety and experiences of harassment, great britain: 16 february to 13 march 2022. https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/perceptionsofpersonalsafetyandexperiencesofharassmentgreatbritain/16februaryto13march2022, 2022. 2, 5, 19

[13] Yasmin B Kofman and Dana Rose Garfin. Home is not always a haven: The domestic violence crisis amid the covid-19 pandemic. *Psychological Trauma: Theory, Research, Practice, and Policy*, 12(S1):S199, 2020. 2

[14] Donna M Gates, Evelyn Fitzwater, and Ursula Meyer. Violence against caregivers in nursing homes: Expected, tolerated, and accepted, 1999. 2

[15] Matthew S. Crow, Jamie A. Snyder, Vaughn J. Crichlow, and John Ortiz Smykla. Community perceptions of police body-worn cameras: The impact of views on fairness, fear, performance, and privacy. *Criminal Justice and Behavior*, 44(4):589–610, 2017. 2

[16] Marion Frost. Health visitors' perceptions of domestic violence: the private nature of the problem. *Journal of Advanced Nursing*, 30(3):589–596, 1999. 3, 12

[17] Sofia Lalanda Frazão, Marília Santos Silva, Pedro Norton, and Teresa Magalhães. Domestic violence against elderly with disability. *Journal of Forensic and Legal Medicine*, 28:19–24, 2014. 3

[18] Kieran Murphy, Sheila Waa, Hussein Jaffer, Agnes Sauter, and Amanda Chan. A literature review of findings in physical elder abuse. *Canadian Association of Radiologists Journal*, 64(1):10–14, 2013. 3

[19] Dorrie E. Rosenblatt, Kyung-Hwan Cho, and Paul W. Durance. Reporting mistreatment of older adults: The role of physicians. *Journal of the American Geriatrics Society*, 44(1):65–70, 1996. 3

[20] Adan Silverio-Murillo, Jose Roberto Balmori de la Miyar, and Lauren Hoehn-Velasco. Families under confinement: Covid-19, domestic violence, and alcohol consumption. *SSRN Electronic Journal*, 2020. 3

[21] Richard R. Peterson and Deirdre Bialo-Padin. Domestic violence is different: The crucial role of evidence collection in domestic violence cases. *Journal of Police Crisis Negotiations*, 12(2):103–121, 2012. 3

[22] Nina J. Westera and Martine B. Powell. Prosecutors' perceptions of how to improve the quality of evidence in domestic violence cases. *Policing and Society*, 27(2):157–172, 2017. 3

[23] Howe Law Firm. Digital evidence in intimate partner abuse cases. https://www.howelawfirm.com/case-types/family-law/domestic-violence/, 2022. 3, 12

[24] Ben Donagh, Caroline Bradbury-Jones, and Julie Taylor. The use of technology to support children and young people experiencing domestic violence and abuse during the covid-19 pandemic: a failure modes and effects analysis. *Journal of Gender-Based Violence*, 6(2):393 – 405, 2022. 4

[25] Diana Freed, Sam Havron, Emily Tseng, Andrea Gallardo, Rahul Chatterjee, Thomas Ristenpart, and Nicola Dell. "is my phone hacked?" analyzing clinical computer security interventions with survivors of intimate partner violence. 3(CSCW), 2019. 4, 12, 134

[26] Hadeel Al-Alosi. Fighting fire with fire: Exploring the potential of technology to help victims combat intimate partner violence. *Aggression and Violent Behavior*, 52:101376, 2020. 4, 6, 12, 17, 134

[27] United Nations. The 17 goals | sustainable development. https://sdgs.un.org/goals, 2023. 4

[28] United Nations. Goal 5 | department of economic and social affairs. https://sdgs.un.org/goals/goal5, 2022. 5

[29] United Nations. Goal 16 | department of economic and social affairs. https://sdgs.un.org/goals/goal16, 2022. 5

[30] Dario Ortega Anderez, Eiman Kanjo, Amna Amnwar, Shane Johnson, and David Lucy. The rise of technology in crime prevention: Opportunities, challenges and practitioners perspectives. *arXiv preprint arXiv:2102.04204*, 2021. 5

[31] Circle of 6. Circle of 6. https://www.circleof6app.com/, 2015 (accessed October 12, 2020). 5, 133

[32] Bright Sky. Bright sky. https://play.google.com/store/apps/details?id=com.newtonmobile.hestia, 2020 (accessed October 14, 2020). 5

[33] Silent Beacon. Silent beacon. https://silentbeacon.com/, 2019 (accessed October 12, 2020). 5, 133, 135

[34] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5):637–646, 2016. 6

[35] Keyan Cao, Yefan Liu, Gongjie Meng, and Qimeng Sun. An overview on edge computing research. *IEEE Access*, 8:85714–85728, 2020. 6

[36] Catherine Hiley. Uk mobile phone statistics 2022. https://www.uswitch.com/mobiles/studies/mobile-statistics/#uk-mobile-phone-user-statistics, 2022. 6, 36, 44

[37] Russell Feldman. Almost a quarter of britons now own one or more smart home devices. https://yougov.co.uk/topics/technology/articles-reports/2018/08/10/almost-quarter-britons-now-own-one-or-more-smart-h, 2018. 6, 17

[38] Mohammed Ali Al-Garadi, Sangmi Kim, Yuting Guo, Elise Warren, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. Natural language model for automatic identification of intimate partner violence reports from twitter. *Array*, 15:100217, 2022. 6, 43

[39] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, 2017. Association for Computational Linguistics. 6, 19

[40] Amira Ghenai and Yelena Mejova. Catching zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on twitter. *CoRR*, abs/1707.03778, 2017. 6

[41] Amrita S. Tulshan and Sudhir Namdeorao Dhage. Survey on virtual assistant: Google assistant, siri, cortana, alexa. In Sabu M. Thampi, Oge Marques, Sri Krishnan, Kuan-Ching Li, Domenico Ciuonzo, and Maheshkumar H. Kolekar, editors, *Advances in Signal Processing and Intelligent Recognition Systems*, pages 190–201, Singapore, 2019. Springer Singapore. 6

[42] Jacopo Biggiogera, George Boateng, Peter Hilpert, Matthew Vowels, Guy Bodenmann, Mona Neysari, Fridtjof Nussbeck, and Tobias Kowatsch. Bert meets liwc: Exploring state-of-the-art language models for predicting communication behavior in couples' conflict interactions. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*,

ICMI '21 Companion, page 385–389, New York, NY, USA, 2021. Association for Computing Machinery. 6, 12, 92

[43] T Chen and Kan Min-Yen. The national university of singapore sms corpus. 2015. 11

[44] Noé Cecillon, Vincent Labatut, Richard Dufour, and Georges Linares. Wac: A corpus of wikipedia conversations for online abuse detection. *arXiv preprint arXiv:2003.06190*, 2020. 11

[45] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015. 12

[46] Eric L Piza, Brandon C Welsh, David P Farrington, and Amanda L Thomas. Cctv surveillance for crime prevention: A 40-year systematic review with meta-analysis. *Criminology & Public Policy*, 18(1):135–159, 2019. 17, 41, 130, 219, 220

[47] Dallas Hill, Christopher D O'Connor, and Andrea Slane. Police use of facial recognition technology: The potential for engaging the public through co-constructed policy-making. *International Journal of Police Science & Management*, 24(3):325–335, 2022. 17

[48] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861, 2020. 18, 62

[49] Varun Mandalapu, Lavanya Elluri, Piyush Vyas, and Nirmalya Roy. Crime prediction using machine learning and deep learning: A systematic review and future directions. *IEEE Access*, 11:60153–60170, 2023. 18

[50] Mark Shaw, Jan Van Dijk, and Wolfgang Rhomberg. Determining trends in global crime and justice: An overview of results from the united nations surveys of crime trends and operations of criminal justice systems. In *Forum on crime and society*, volume 3, pages 35–63. Citeseer, 2003. 18

[51] Hugh Downing. The emergence of global positioning satellite (gps) systems in correctional applications. *Corrections Today*, 68(6):42, 2006. 18, 229

[52] Anas Basalamah. Sensing the crowds using bluetooth low energy tags. *IEEE access*, 4:4225–4233, 2016. 18

[53] Jyoti Belur, Amy Thornton, Lisa Tompson, Matthew Manning, Aiden Sidebottom, and Kate Bowers. A systematic review of the effectiveness of the electronic monitoring of offenders. *Journal of Criminal Justice*, 68:101686, 2020. 18, 227

[54] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F Churchill, and Sunny Consolvo. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2189–2201. ACM, 2017. 18, 43

[55] Hitesh Kumar Sharma, K Kshitiz, et al. Nlp and machine learning techniques for detecting insulting comments on social networking platforms. In *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pages 265–272. IEEE, 2018. 18, 22, 56

[56] Veeramanikandan, Suresh Sankaranarayanan, Joel J.P.C. Rodrigues, Vijayan Sugumaran, and Sergei Kozlov. Data flow and distributed deep neural network based low latency iot-edge computation model for big data environment. *Engineering Applications of Artificial Intelligence*, 94:103785, 2020. 18

[57] Md Saroar Jahan and Mourad Oussalah. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232, 2023. 18

[58] Ying Chen. Detecting offensive language in social medias for protection of adolescent online safety. 2011. 18, 31, 72

[59] Rasha Zieni, Luisa Massari, and Maria Carla Calzarossa. Phishing or not phishing? a survey on the detection of phishing websites. *IEEE Access*, 11:18499–18519, 2023. 18

[60] Emad E. Abdallah Jamil R. Alzghoul and Abdel hafiz S. Al-khawaldeh. Fraud in online classified ads: Strategies, risks, and detection methods: A survey. *Journal of Applied Security Research*, 19(1):45–69, 2024. 18

[61] Shivam B. Parikh, Saurin R. Khedia, and Pradeep K. Atrey. A framework to detect fake tweet images on social media. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 104–110, 2019. 18

[62] Meghan Elkin. Crime in england and wales: Year ending march 2021. https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingmarch2021, 2021. 19

[63] Police UK. Advice and crime prevention | police.uk. https://www.police.uk/cp/crime-prevention/, 2023. 19

[64] Metropolitan Police. Mark your property to deter burglars | metropolitan police. https://www.met.police.uk/cp/crime-prevention/protect-home-crime/mark-your-property/, 2023. 19

[65] Paul Ekblom. *Technology, Opportunity, Crime and Crime Prevention: Current and Evolutionary Perspectives*, pages 319–343. Springer International Publishing, Cham, 2017. 19

[66] Mario E. Ninaquispe Soto, Yasmina Riega-Virú, and Juan C. Oruna Lara. Technology and crime prevention: A systematic review of literature. In *2021 Fifth World Conference on Smart Trends in Systems Security and Sustainability (WorldS4)*, pages 275–283, 2021. 19

[67] Temidayo James Aransiola and Vania Ceccato. The role of modern technology in rural situational crime prevention: A review of the literature. *Rural crime prevention*, pages 58–72, 2020. 20

[68] Selma Dilek, Hüseyin Çakir, and Mustafa Aydin. Applications of artificial intelligence techniques to combating cyber crimes: A review. *CoRR*, abs/1502.03552, 2015. 20

[69] Samuel J. Stratton. Literature reviews: Methods and applications. *Prehospital and Disaster Medicine*, 34(4):347–349, 2019. 20, 21

[70] Cynthia L. Russell. An overview of the integrative research review. *Progress in Transplantation*, 15(1):8–13, 2005. 21

[71] Filip Bacalu et al. Digital policing tools as social control technologies: data-driven predictive algorithms, automated facial recognition surveillance, and law enforcement biometrics. *Analysis and Metaphysics*, (20):74–88, 2021. 21, 222

[72] Weijun Qin, Jiadi Zhang, Bo Li, and Limin Sun. Discovering human presence activities with smartphones using nonintrusive wi-fi sniffer sensors: the big data prospective. *International Journal of Distributed Sensor Networks*, 9(12):927940, 2013. 21

[73] Yizhao Ni, Drew Barzman, Alycia Bachtel, Marcus Griffey, Alexander Osborn, and Michael Sorter. Finding warning markers: leveraging natural language processing and machine learning technologies to detect risk of school violence. *International journal of medical informatics*, 139:104137, 2020. 21, 31, 33

[74] Ying-Lung Lin, Tenge-Yang Chen, and Liang-Chih Yu. Using machine learning to assist crime prevention. In *2017 6th IIAI international congress on advanced applied informatics (IIAI-AAI)*, pages 1029–1030. IEEE, 2017. 21, 22, 34

[75] Adrienne Yapo and Joseph Weiss. Ethical implications of bias in machine learning. 2018. 21

[76] Nicol Turner Lee. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, 16(3):252–260, 2018. 21

[77] Issam El Naqa and Martin J Murphy. What is machine learning? In *machine learning in radiation oncology*, pages 3–11. Springer, 2015. 22

[78] Meysam Vakili, Mohammad Ghamsari, and Masoumeh Rezaei. Performance analysis and comparison of machine and deep learning algorithms for iot data classification. *arXiv preprint arXiv:2001.09636*, 2020. 22

[79] Sanika Tanmay Ratnaparkhi, Aamani Tandasi, and Shipra Saraswat. Face detection and recognition for criminal identification system. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 773–777. IEEE, 2021. 22

[80] Chedia Dhaoui, Cynthia M Webster, and Lay Peng Tan. Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing*, 2017. 22

[81] Wilpen L Gorr and YongJei Lee. Early warning system for temporary crime hot spots. *Journal of Quantitative Criminology*, 31(1):25–47, 2015. 22

[82] Amitha Mathew, P Amudha, and S Sivakumari. Deep learning techniques: an overview. *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020*, pages 599–608, 2021. 22

[83] Nischal Sanil, V Rakesh, Rishab Mallapur, Mohammed Riyaz Ahmed, et al. Deep learning techniques for obstacle detection and avoidance in driverless cars. In *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*, pages 1–4. IEEE, 2020. 22

[84] C Harris. Police and soft technology: How information technology contributes to police decision making. *The new technology of crime, law and social control*, pages 153–183, 2007. 22

[85] Noushin Hajarolasvadi and Hasan Demirel. Deep facial emotion recognition in video using eigenframes. *IET Image Processing*, 2020. 22, 56

[86] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and trends® in signal processing*, 7(3–4):197–387, 2014. 22

[87] Sweta Agrawal and Amit Awekar. Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, pages 141–153. Springer, 2018. 23

[88] Chang-Soo Sung and Joo Yeon Park. Design of an intelligent video surveillance system for crime prevention: applying deep learning technology. *Multimedia Tools and Applications*, 80(26):34297–34309, 2021. 23, 53

[89] Sun-Chong Wang. Artificial neural network. In *Interdisciplinary computing in java programming*, pages 81–100. Springer, 2003. 23

[90] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11):e00938, 2018. 25

[91] Andreas M Olligschlaeger. Artificial neural networks and crime mapping. *Crime mapping and crime prevention*, 1:313, 1997. 25, 43

[92] Syahid Anuar, Ali Selamat, and Roselina Sallehuddin. Hybrid artificial neural network with artificial bee colony algorithm for crime classification. In *Computational Intelligence in Information Systems*, pages 31–40. Springer, 2015. 25

[93] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015. 26

[94] Qian Huang and Kangli Hao. Development of cnn-based visual recognition air conditioner for smart buildings. *J. Inf. Technol. Constr.*, 25:361–373, 2020. 26

[95] C. Hema and Fausto Pedro Garcia Marquez. Emotional speech recognition using cnn and deep learning techniques. *Applied Acoustics*, 211:109492, 2023. 26

[96] Wei Wang and Jianxun Gang. Application of convolutional neural network in natural language processing. In *2018 International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pages 64–70. IEEE, 2018. 26

[97] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018. 26, 27

[98] Xin Yao, Xiaoran Shi, and Feng Zhou. Human activities classification based on complex-value convolutional neural network. *IEEE Sensors Journal*, 20(13):7169–7180, 2020. 27

[99] Guillermo A Martínez-Mascorro, José R Abreu-Pederzini, José C Ortiz-Bayliss, and Hugo Terashima-Marín. Suspicious behavior detection on shoplifting cases for crime prevention by using 3d convolutional neural networks. *arXiv preprint arXiv:2005.02142*, 2020. 27

[100] Gennady Andrienko, Natalia Andrienko, Urska Demsar, Doris Dransch, Jason Dykes, Sara Irina Fabrikant, Mikael Jern, Menno-Jan Kraak, Heidrun Schumann, and Christian Tominski. Space, time and visual analytics. *International journal of geographical information science*, 24(10):1577–1600, 2010. 28

[101] Zhicheng Shi and Lilian SC Pun-Cheng. Spatiotemporal data clustering: a survey of methods. *ISPRS international journal of geo-information*, 8(3):112, 2019. 28

[102] Umair Muneer Butt, Sukumar Letchmunan, Fadratul Hafinaz Hassan, Mubashir Ali, Anees Baqir, and Hafiz Husnain Raza Sherazi. Spatio-temporal crime hotspot detection and prediction: a systematic literature review. *IEEE Access*, 8:166553–166574, 2020. 28

[103] Senzhang Wang, Jiannong Cao, and Philip Yu. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, 2020. 28

[104] Ying-Lung Lin, Meng-Feng Yen, and Liang-Chih Yu. Grid-based crime prediction using geographical features. *ISPRS International Journal of Geo-Information*, 7(8):298, 2018. 28

[105] Plamen P Angelov and Xiaowei Gu. *Empirical approach to machine learning*. Springer, 2019. 29

[106] Anthony A Braga, Brandon S Turchan, Andrew V Papachristos, and David M Hureau. Hot spots policing and crime reduction: An update of

an ongoing systematic review and meta-analysis. *Journal of experimental criminology*, 15(3):289–311, 2019. 29

[107] Hoang Nguyen, Xuan-Nam Bui, Quang-Hieu Tran, Pham Van Hoa, Dinh-An Nguyen, Le Thi Thu Hoa, Qui-Thao Le, Ngoc-Hoan Do, Tran Dinh Bao, Hoang-Bac Bui, et al. A comparative study of empirical and ensemble machine learning algorithms in predicting air over-pressure in open-pit coal mine. *Acta Geophysica*, 68(2):325–336, 2020. 29

[108] Matthew Manning, Gabriel TW Wong, Timothy Graham, Thilina Ranbaduge, Peter Christen, Kerry Taylor, Richard Wortley, Toni Makkai, and Pierre Skorich. Towards a 'smart'cost–benefit tool: using machine learning to predict the costs of criminal justice policy interventions. *Crime Science*, 7(1):1–13, 2018. 29

[109] Celestine Iwendi, Gautam Srivastava, Suleman Khan, and Praveen Kumar Reddy Maddikunta. Cyberbullying detection solutions based on deep learning architectures. *Multimedia Systems*, 29(3):1839–1852, 2023. 30

[110] Monisha Kanakaraj and Ram Mohana Reddy Guddeti. Nlp based sentiment analysis on twitter data using ensemble classifiers. In *2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN)*, pages 1–5, 2015. 30

[111] E Rutger Leukfeldt. Cybercrime and social ties. *Trends in organized crime*, 17(4):231–249, 2014. 30

[112] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 30, 74, 75, 76

[113] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In Hocine Cherifi, Sabrina Gaito, José Fernendo Mendes, Esteban Moro, and Luis Mateus Rocha, editors, *Complex Networks and Their Applications VIII*, pages 928–940, Cham, 2020. Springer International Publishing. 30

[114] Maryam Heidari, Samira Zad, Parisa Hajibabaee, Masoud Malekzadeh, SeyyedPooya HekmatiAthar, Ozlem Uzuner, and James H Jones. Bert model for fake news detection based on social bot activities in the covid-19 pandemic. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pages 0103–0109, 2021. 30

[115] Nickolas M Jones, Sean P Wojcik, Josiah Sweeting, and Roxane Cohen Silver. Tweeting negative emotion: An investigation of twitter data in the aftermath of violence on college campuses. *Psychological methods*, 21(4):526, 2016. 30

[116] Rodrigo Augusto Silva Dos Santos et al. *Robust Noise-Based Attacks against Audio Event Detection Systems*. PhD thesis, 2022. 30, 31

[117] Francisco J Pérez, Victor J Garrido, Alberto García, Marcelo Zambrano, Rafał Kozik, Michał Choraś, Dirk Mühlenberg, Dirk Pallmer, and Wilmuth Müller. Multimedia analysis platform for crime prevention and investigation. *Multimedia Tools and Applications*, 80(15):23681–23700, 2021. 30, 226

[118] Muhammad Zidny Naf'an, Alhamda Adisoka Bimantara, Afiatari Larasati, Ezar Mega Risondang, and Novanda Alim Setya Nugraha. Sentiment analysis of cyberbullying on instagram user comments. *Journal of Data Science and Its Applications*, 2(1):38–48, 2019. 31, 72

[119] Abayomi Bello, Sin-Chun Ng, and Man-Fai Leung. A bert framework to sentiment analysis of tweets. *Sensors*, 23(1), 2023. 31

[120] Doaa Mohey El-Din Mohamed Hussein. A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4):330–338, 2018. 31, 72

[121] Iti Chaturvedi, Erik Cambria, Roy E. Welsch, and Francisco Herrera. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, 44:65–77, 2018. 31, 72

[122] Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*, pages 1–6, 2016. 31, 72

[123] Feiyang Chen, Ziqian Luo, Yanyan Xu, and Dengfeng Ke. Complementary fusion of multi-features and multi-modalities in sentiment analysis. *arXiv preprint arXiv:1904.08138*, 2019. 31, 72

[124] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE, 2012. 32, 62, 72, 96

[125] Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80, 2017. 32, 72

[126] Jeremy Crump. What are the police doing on twitter? social media, the police and the public. *Policy & Internet*, 3(4):1–27, 2011. 32

[127] Cristina Kadar, Yiea-Funk Te, Raquel Rosés Brüngger, and Irena Pletikosa Cvijikj. Digital neighborhood watch: To share or not to share? In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, page 2148–2155, New York, NY, USA, 2016. Association for Computing Machinery. 32

[128] Aarti Israni, Sheena Erete, and Che L Smith. Snitches, trolls, and social norms: Unpacking perceptions of social media use for crime prevention. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1193–1209, 2017. 32

[129] Konstantinos Domdouzis, Babak Akhgar, Simon Andrews, Helen Gibson, and Laurence Hirsch. A social media and crowdsourcing data mining system for crime prevention during and post-crisis situations. *Journal of Systems and Information Technology*, 2016. 32

[130] Sunmisola Eniola Peters and Usman Adekunle Ojedokun. Social media utilization for policing and crime prevention in lagos, nigeria. *Journal of social, behavioral, and health sciences*, 13(1):11, 2019. 33

[131] Panote Siriaraya, Yihong Zhang, Yuanyuan Wang, Yukiko Kawai, Mohit Mittal, Péter Jeszenszky, and Adam Jatowt. Witnessing crime through tweets: a crime investigation tool based on social media. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 568–571, 2019. 33

[132] Naomi Smith. Social media, rural communities and crime prevention. In *Rural Crime Prevention*, pages 73–83. Routledge, 2020. 33

[133] Arielle Stephenson, Gal Avisar Cohen, Yitzchak Beller, and Dov Greenbaum. Suicide prevention technologies and social media platforms: Legal, social and ethical implications. *Social and Ethical Implications (August 18, 2021)*, 2021. 33

[134] Joie Ann W. Maghanoy. Crime mapping report mobile application using gis. In *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, pages 247–251, 2017. 33, 35

[135] JavedAnjum Sheikh, Iqra Shafique, Madiha Sharif, Syeda Ambreen Zahra, and Tuba Farid. Ist: Role of gis in crime mapping and analysis. In *2017 International Conference on Communication Technologies (ComTech)*, pages 126–131, 2017. 33, 35

[136] Sarah Barns. Smart cities and urban data platforms: Designing interfaces for smart governance. *City, culture and society*, 12:5–12, 2018. 33

[137] Piyush Kumar, Akhil Kumar Jha, and B. Balamurugan. Dashboard for crime analytics using r shiny. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 783–789, 2022. 34

[138] Gabriel Orsini, Dirk Bade, and Winfried Lamersdorf. Cloudaware: A context-adaptive middleware for mobile edge and cloud computing appli-

cations. In *2016 IEEE 1st International Workshops on Foundations and Applications of Self* Systems (FAS* W)*, pages 216–221. IEEE, 2016. 34

[139] Sinung Suakanto, Suhono H Supangkat, Roberd Saragih, et al. Smart city dashboard for integrating various data of sensor networks. In *International Conference on ICT for Smart Society*, pages 1–5. IEEE, 2013. 34

[140] Oliver Hutt, Kate Bowers, Shane Johnson, and Toby Davies. Data and evidence challenges facing place-based policing. *Policing: An International Journal*, 41(3):339–351, 2018. 34

[141] Rachel Boba Santos. *Crime analysis with crime mapping*. Sage publications, 2016. 34

[142] Shane D Johnson. A brief history of the analysis of crime concentration. *European Journal of Applied Mathematics*, 21(4-5):349–370, 2010. 34

[143] Adam A Abbas, Abubakar U Alhaji, and Kadini J Bitrus. Locational analysis of crime in gombe metropolis, nigeria. 2017. 34

[144] Anthony A Braga, Martin A Andresen, and Brian Lawton. The law of crime concentration at places: Editors' introduction, 2017. 34

[145] Simon Williams and Timothy Coupe. Frequency vs. length of hot spots patrols: a randomised controlled trial. *Cambridge Journal of Evidence-Based Policing*, 1(1):5–21, 2017. 34

[146] Germain Garcia-Zanabria, Erick Gomez-Nieto, Jaqueline Silveira, Jorge Poco, Marcelo Nery, Sergio Adorno, and Luis G. Nonato. Mirante: A visualization tool for analyzing urban crimes. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 148–155, 2020. 34

[147] Guiyun Zhou, Jiayuan Lin, and Wenfeng Zheng. A web-based geographical information system for crime mapping and decision support. In *2012 International Conference on Computational Problem-Solving (ICCP)*, pages 147–150, 2012. 35

[148] Zaheer Khan and Saad Liaquat Kiani. A cloud-based architecture for citizen services in smart cities. In *2012 IEEE Fifth International Conference on Utility and Cloud Computing*, pages 315–320, 2012. 35

[149] J Byrne. The new generation of concentrated community supervision strategies: Focusing resources on high risk offenders, times, and places. *A Report for the Public Safety Performance Project, the Pew Charitable Trusts: Washington, DC*, 2009. 35

[150] James Byrne and Gary Marx. Technological innovations in crime prevention and policing. a review of the research on implementation and impact. *Journal of Police Studies*, 20(3):17–40, 2011. 35, 220

[151] Christopher Soghoian. The law enforcement surveillance reporting gap. *Available at SSRN 1806628*, 2011. 36

[152] Red Panic Button. Red panic button. https://apps.apple.com/us/app/red-panic-button/id422029296, 2020 (accessed October 14, 2020). 36

[153] Apple. icloud - find my - apple (uk). https://www.apple.com/uk/icloud/find-my/, 2023. 36

[154] Life360 Inc. Homepage - life360, 2023. https://www.life360.com/intl/. 39

[155] Dario Ortega Anderez, Ahmad Lotfi, and Amir Pourabdollah. Eating and drinking gesture spotting and recognition using a novel adaptive segmentation technique and a gesture discrepancy measure. *Expert Systems with Applications*, 140:112888, 2020. 41, 47, 132

[156] Clifford S Fishman. The interception of communications without a court order: Title iii, consent, and the expectation of privacy. . *John's L. Rev.*, 51:41, 1976. 41, 222

[157] Jerry H Ratcliffe, Matthew Lattanzio, George Kikuchi, and Kevin Thomas. A partially randomized field experiment on the effect of an acoustic gunshot detection system on police incident reports. *Journal of Experimental Criminology*, 15(1):67–76, 2019. 41

[158] Pierre Laffitte, Yun Wang, David Sodoyer, and Laurent Girin. Assessing the performances of different neural network architectures for the detection of screams and shouts in public transportation. *Expert Systems With Applications*, 117:29–41, 2019. 41

[159] Jianfeng Chen, Alvin Harvey Kam, Jianmin Zhang, Ning Liu, and Louis Shue. Bathroom activity monitoring based on sound. In *International Conference on Pervasive Computing*, pages 47–61. Springer, 2005. 41

[160] Eamonn O'Neill, Vassilis Kostakos, Tim Kindberg, Alan Penn, Danaë Stanton Fraser, Tim Jones, et al. Instrumenting the city: Developing methods for observing and understanding the digital cityscape. In *International Conference on Ubiquitous Computing*, pages 315–332. Springer, 2006. 42, 232

[161] Zhenyu Chen, Yiqiang Chen, Xingyu Gao, Shuangquan Wang, Lisha Hu, Chenggang Clarence Yan, Nicholas D Lane, and Chunyan Miao. Unobtrusive sensing incremental social contexts using fuzzy class incremental learning. In *2015 IEEE International Conference on Data Mining*, pages 71–80. IEEE, 2015. 42, 232

[162] Mahasak Ketcham, Thittaporn Ganokratanaa, and Sriphagaarucht Srinhichaarnun. The intruder detection system for rapid transit using cctv surveillance based on histogram shapes. In *2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–6. IEEE, 2014. 43

[163] John Eck, Spencer Chainey, James Cameron, and Ronald Wilson. Mapping crime: Understanding hotspots. 2005. 44

[164] Hadi Habibzadeh, Brian H Nussbaum, Fazel Anjomshoa, Burak Kantarci, and Tolga Soyata. A survey on cybersecurity, data privacy, and policy issues in cyber-physical system deployments in smart cities. *Sustainable Cities and Society*, 2019. 45

[165] Rida Khatoun and Sherali Zeadally. Cybersecurity and privacy solutions in smart cities. *IEEE Communications Magazine*, 55(3):51–59, 2017. 45

[166] Nusrat J Shoumy, Li-Minn Ang, Kah Phooi Seng, DM Motiur Rahaman, and Tanveer Zia. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *Journal of Network and Computer Applications*, 149:102447, 2020. 47

[167] Jordan J Bird, Elizabeth Wanner, Anikó Ekárt, and Diego R Faria. Accent classification in human speech biometrics for native and non-native english speakers. In *PETRA*, pages 554–560, 2019. 47

[168] Eiman Kanjo, Eman MG Younis, and Chee Siang Ang. Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion*, 49:46–56, 2019. 47

[169] Dario Ortega-Anderez, Ahmad Lotfi, Caroline Langensiepen, and Kofi Appiah. A multi-level refinement approach towards the classification of quotidian activities using accelerometer data. *Journal of Ambient Intelligence and Humanized Computing*, 10(11):4319–4330, 2019. 47

[170] John L. Gustafson. *Moore's Law*, pages 1177–1184. Springer US, Boston, MA, 2011. 47

[171] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "a stalker's paradise": How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery. 48

[172] Delanie Woodlock. The abuse of technology in domestic violence and stalking. *Violence Against Women*, 23(5):584–602, 2017. 48

[173] Delanie Woodlock, Mandy McKenzie, Deborah Western, and Bridget Harris. Technology as a weapon in domestic violence: Responding to digital coercive control. *Australian Social Work*, 73(3):368–380, 2020. 48

[174] Julio Suarez-Paez, Mayra Salcedo-Gonzalez, Alfonso Climente, Manuel Esteve, Jon Ander Gómez, Carlos Enrique Palau, and Israel Pérez-Llopis. A novel low processing time system for criminal activities detection applied

to command and control citizen security centers. *Information*, 10(12):365, 2019. 50

[175] Dingqi Yang, Terence Heaney, Alberto Tonon, Leye Wang, and Philippe Cudré-Mauroux. Crimetelescope: crime hotspot prediction based on urban and social media data fusion. *World Wide Web*, 21(5):1323–1347, 2018. 52

[176] Hyeon-Woo Kang and Hang-Bong Kang. Prediction of crime occurrence from multi-modal data using deep learning. *PloS one*, 12(4):e0176244, 2017. 52

[177] Jianhu Zheng and Jinshuan Peng. A novel pedestrian detection algorithm based on data fusion of face images. *International Journal of Distributed Sensor Networks*, 15(5):1550147719845276, 2019. 52

[178] David L Weisburd and Tom McEwen. Introduction: Crime mapping and crime prevention. *Available at SSRN 2629850*, 2015. 53

[179] Roy F. Baumeister and Brad J. Bushman. *Emotions and Aggressiveness*, pages 479–493. Springer Netherlands, Dordrecht, 2003. 55

[180] Terry Allen, Shannon A Novak, and Lawrence L Bench. Patterns of injuries: accident or abuse. *Violence Against Women*, 13(8):802–816, 2007. 55

[181] Anjali Bhavan, Pankaj Chauhan, Rajiv Ratn Shah, et al. Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems*, 184:104886, 2019. 56

[182] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017. 56

[183] Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In *Interspeech*, pages 2803–2807, 2019. 56, 96

[184] Bagus Tris Atmaja and Masato Akagi. Speech emotion recognition based on speech segment using lstm with attention model. In *2019 IEEE International Conference on Signals and Systems (ICSigSys)*, pages 40–44. IEEE, 2019. 56, 96

[185] Zhengyin Du, Suowei Wu, Di Huang, Weixin Li, and Yunhong Wang. Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition. *IEEE Transactions on Affective Computing*, 2019. 56

[186] Min Hu, Haowen Wang, Xiaohua Wang, Juan Yang, and Ronggui Wang. Video facial emotion recognition based on local enhanced motion history image and cnn-ctslstm networks. *Journal of Visual Communication and Image Representation*, 59:176–185, 2019. 56

[187] Erdenebileg Batbaatar, Meijing Li, and Keun Ho Ryu. Semantic-emotion neural network for emotion recognition from text. *IEEE Access*, 7:111866–111878, 2019. 56

[188] Cheng-Ta Yang and Yi-Ling Chen. Dacnn: Dynamic weighted attention with multi-channel convolutional neural network for emotion recognition. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pages 316–321. IEEE, 2020. 56

[189] Flor Miriam Plaza-del Arco, M Teresa Martín-Valdivia, L Alfonso Ureña-López, and Ruslan Mitkov. Improved emotion recognition in spanish social media through incorporation of lexical knowledge. *Future Generation Computer Systems*, 110:1000–1008, 2020. 56

[190] Aravind K Joshi. Natural language processing. *Science*, 253(5025):1242–1249, 1991. 56

[191] Zewdie Mossie and Jenq-Haur Wang. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087, 2020. 56

[192] Theodoros Giannakopoulos, Dimitrios Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis. Violence content classification using audio fea-

tures. In Grigoris Antoniou, George Potamias, Costas Spyropoulos, and Dimitris Plexousakis, editors, *Advances in Artificial Intelligence*, pages 502–507, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. 56

[193] Herbert A. Simon. The science of design: Creating the artificial. *Design Issues*, 4(1/2):67–82, 1988. 57, 58

[194] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77, 2007. 57

[195] Richard Baskerville. What design science is not. *European Journal of Information Systems*, 17(5):441–443, 2008. 57

[196] Kieran Woodward, Eiman Kanjo, Dario Ortega Anderez, Amna Anwar, Thomas Johnson, and John Hunt. Digitalppe: low cost wearable that acts as a social distancingreminder and contact tracer. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 758–759, 2020. 60

[197] Dario Ortega Anderez, Eiman Kanjo, Ganna Pogrebna, Omprakash Kaiwartya, Shane D Johnson, and John Alan Hunt. A covid-19-based modified epidemiological model and technological approaches to help vulnerable individuals emerge from the lockdown in the uk. *Sensors*, 20(17):4967, 2020. 60

[198] Sara Owsley Sood, Judd Antin, and Elizabeth Churchill. Using crowdsourcing to improve profanity detection. In *2012 AAAI Spring Symposium Series*, 2012. 62

[199] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152, 2019. 62

[200] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016. 62

[201] Gray Atkins. Eastenders. gray's torture abuse of chantelle atkins full story, 2020. 63

[202] R Regin, S Suman Rajest, T Shynu, et al. An automated conversation system using natural language processing (nlp) chatbot in python. *Central Asian Journal of Medical and Natural Science*, 3(4):314–336, 2022. 71

[203] Sunil Malviya, Arvind Kumar Tiwari, Rajeev Srivastava, and Vipin Tiwari. Machine learning techniques for sentiment analysis: A review. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, 12(02):72–78, 2020. 71

[204] Basant Agarwal, Namita Mittal, Basant Agarwal, and Namita Mittal. Machine learning approach for sentiment analysis. *Prominent feature extraction for sentiment analysis*, pages 21–45, 2016. 71

[205] Ike Vayansky and Sathish AP Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020. 71

[206] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007. 71

[207] Jia Xue, Junxiang Chen, and Richard Gelles. Using data mining techniques to examine domestic violence topics on twitter. *Violence and gender*, 6(2):105–114, 2019. 71

[208] Behrang Mohit. Named entity recognition. *Natural language processing of semitic languages*, pages 221–245, 2014. 71

[209] Ricardo Martins, Marco Gomes, Jose Joao Almeida, Paulo Novais, and Pedro Henriques. Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 61–66. IEEE, 2018. 71

[210] Amgad Muneer and Suliman Mohamed Fati. A comparative analysis of machine learning techniques for cyberbullying detection on twitter. *Future Internet*, 12(11):187, 2020. 71

[211] Matthew Edwards, Awais Rashid, and Paul Rayson. A systematic survey of online data mining technology intended for law enforcement. *ACM Computing Surveys (CSUR)*, 48(1):1–54, 2015. 71

[212] A Gharat, H Tandel, and K Bagade. Natural language processing theory applications and difficulties. *IJTSRD*, 3(6):501–503, 2019. 71

[213] Alvaro Ortigosa, José M Martín, and Rosa M Carro. Sentiment analysis in facebook and its application to e-learning. *Computers in human behavior*, 31:527–541, 2014. 72

[214] Aliaksei Severyn and Alessandro Moschitti. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 959–962, 2015. 72

[215] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26, 2012. 72

[216] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015. 72

[217] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019. 72, 74

[218] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018. 74

[219] Yuanbin Qu, Peihan Liu, Wei Song, Lizhen Liu, and Miaomiao Cheng. A text generation and prediction system: Pre-training on new corpora using bert and gpt-2. In *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pages 323–326. IEEE, 2020. 74

[220] Hind S Alatawi, Areej M Alhothali, and Kawthar M Moria. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374, 2021. 74

[221] mekarahul. What are self-attention models? https://medium.com/@mekarahul/what-are-self-attention-models-69fb59f6b5f8, Accessed on 19th July 2023. 76

[222] Daniel Jurafsky and James H. Martin. Speech and language processing (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/3.pdf, Accessed on 19th July 2023. 77

[223] Divyanshu. Lstm and its equations. https://medium.com/@divyanshu132/lstm-and-its-equations-5ee9246d04af, Accessed on 19th April 2023. 78

[224] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. 2015. 79, 80, 81

[225] LIWC. Liwc - liwc analysis. https://www.liwc.app/help/liwc, 2023. 81

[226] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720, 2018. 91, 93, 127, 166, 167, 168

[227] James R. Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F. Quatieri. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, page 11–18, New York, NY, USA, 2016. Association for Computing Machinery. 91, 127

[228] Arjumand Younus and M. Atif Qureshi. Combining bert with contextual linguistic features for identification of propaganda spans in news articles. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 5864–5866, 2020. 92

[229] Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen. Sentiment extraction from natural audio streams. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8485–8489. IEEE, 2013. 96

[230] S Maghilnan and M Rajesh Kumar. Sentiment analysis on speaker specific speech data. In *2017 International Conference on Intelligent Computing and Control (I2C2)*, pages 1–5. IEEE, 2017. 96

[231] Ziqian Luo, Hua Xu, and Feiyang Chen. Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural network. In *AffCon@ AAAI*, 2019. 96

[232] Bryan Li, Dimitrios Dimitriadis, and Andreas Stolcke. Acoustic and lexical sentiment analysis for customer service calls. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5876–5880. IEEE, 2019. 96

[233] Archana L Rane and Ankita R Kshatriya. Audio opinion mining and sentiment analysis of customer product or services reviews. In *ICDSMLA 2019*, pages 282–293. Springer, 2020. 96

[234] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. Speech emotion recognition using spectrogram & phoneme embedding. In *Interspeech*, pages 3688–3692, 2018. 96

[235] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 96

[236] Filip Povolny, Pavel Matejka, Michal Hradis, Anna Popková, Lubomír Otrusina, Pavel Smrz, Ian Wood, Cecile Robin, and Lori Lamel. Multimodal emotion recognition for avec 2016 challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 75–82, 2016. 96, 113

[237] Gaurav Sahu. Multimodal speech emotion recognition and ambiguity resolution. *arXiv preprint arXiv:1904.06022*, 2019. 96

[238] Moisés Henrique Ramos Pereira, Flávio Luis Cardeal Pádua, Adriano César Machado Pereira, Fabrício Benevenuto, and Daniel Hasan Dalip. Fusing audio, textual, and visual features for sentiment analysis of news videos. In *Tenth International AAAI Conference on Web and Social Media*, 2016. 96

[239] Mauricio Perez-Sosa, A. Baki Kocaballi, Jorge Quiroz, Uthayasanker Wijesundara, Luiz G. Hafemann, and Luciano Oliveira. Audiovisual emotion recognition for detection of domestic violence: A feasibility study. *IEEE Access*, 8:188658–188667, 2020. 98

[240] Tharindu H De Silva, Kaveesha K Jayawardena, Anil Fernando, Samadhi T Ekanayake, Udaya Yapa, Asanka W Jayawardena, and Roshan Jayasekara. Recognising domestic violence from audio data using deep learning. *arXiv preprint arXiv:2006.02962*, 2020. 98

[241] Donghee Kang, Jacob M Whitehill, Thomas A Lasko, Barbara P Buttenfield, and Jack A Gorman. Textual analysis of domestic violence police reports. *Journal of Interpersonal Violence*, 28(8):1659–1680, 2013. 98

[242] Saif Mohammad, Muhammad Imran, Prasenjit Mitra, and Svetlana Kiritchenko. Empirical study of online behaviours of radicalised youth. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1650–1661, 2016. 99

[243] Chia-Ping Chen. Discrete-time signals and systems. 1983. 107

[244] Haytham M. Fayek. Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients (mfccs) and what's in-between, 2016. 107, 108

[245] Jar-Ferr Yang and Fu-Kun Chen. Recursive discrete fourier transform with unified iir filter structures. *Signal Processing*, 82(1):31–41, 2002. 107

[246] Maxim Raginsky. Lecture xi: The fast fourier transform (fft) algorithm. https://maxim.ece.illinois.edu/teaching/fall08/lec11.pdf, 2008. 108

[247] J. Makhoul. A fast cosine transform in one and two dimensions. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(1):27–34, 1980. 108

[248] Genevieve Lam, Huang Dongyan, and Weisi Lin. Context-aware deep learning for multi-modal depression detection. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3946–3950, 2019. 110, 127, 166, 168

[249] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2020. 111

[250] Hervé Bredin and Antoine Laurent. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*, 2021. 111

[251] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017. 113

[252] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982, 2013. 113

[253] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53, 2013. 113

[254] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544, 2015. 113

[255] Blesson Varghese, Nan Wang, Sakil Barbhuiya, Peter Kilpatrick, and Dimitrios S. Nikolopoulos. Challenges and opportunities in edge computing. In *2016 IEEE International Conference on Smart Cloud (SmartCloud)*, pages 20–26, 2016. 130

[256] Dario Ortega Anderez, Eiman Kanjo, Amna Anwar, Shane Johnson, and David Lucy. The rise of technology in crime prevention: Opportunities, challenges and practitioners perspectives, 2021. 130

[257] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5):637–646, 2016. 130, 131, 132, 160

[258] Michael Losavio, Adel Elmaghraby, and Antonio Losavio. Ubiquitous networks, ubiquitous sensors: Issues of security, reliability and privacy in the internet of things. In *International Symposium on Ubiquitous Networking*, pages 331–343. Springer, 2018. 130, 227

[259] Sudhir Chitnis, Neha Deshpande, Arvind Shaligram, et al. An investigative study for smart home security: Issues, challenges and countermeasures. *Wireless Sensor Network*, 8(04):61, 2016. 130, 225

[260] Sumit Shah, Fenye Bao, Chang-Tien Lu, and Ing-Ray Chen. Crowdsafe: Crowd sourcing of crime incidents and safe routing on mobile devices. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, page 521–524, New York, NY, USA, 2011. Association for Computing Machinery. 131

[261] Lamiaa A. Elrefaei, Alaa Alharthi, Huda Alamoudi, Shatha Almutairi, and Fatima Al-rammah. Real-time face detection and tracking on mobile phones for criminal detection. In *2017 2nd International Conference on Anti-Cyber Crimes (ICACC)*, pages 75–80, 2017. 131

[262] Mark A. Wood, Stuart Ross, and Diana Johns. Primary crime prevention apps: A typology and scoping review. *Trauma, Violence, & Abuse*, 23(4):1093–1110, 2022. PMID: 33478344. 131, 160

[263] Latif U. Khan, Ibrar Yaqoob, Nguyen H. Tran, S. M. Ahsan Kazmi, Tri Nguyen Dang, and Choong Seon Hong. Edge-computing-enabled smart cities: A comprehensive survey. *IEEE Internet of Things Journal*, 7(10):10200–10232, 2020. 132

[264] Yaser Jararweh, Ahmad Doulat, Omar AlQudah, Ejaz Ahmed, Mahmoud Al-Ayyoub, and Elhadj Benkhelifa. The future of mobile cloud computing: Integrating cloudlets and mobile edge computing. In *2016 23rd International Conference on Telecommunications (ICT)*, pages 1–5, 2016. 132

[265] Dhruv Chand, Sunil Nayak, Karthik S. Bhat, Shivani Parikh, Yuvraj Singh, and Amita Ajith Kamath. A mobile application for women's safety: Wosapp. In *TENCON 2015 - 2015 IEEE Region 10 Conference*, pages 1–5, 2015. 133, 134, 135

[266] tflite_flutter: Flutter package. https://pub.dev/packages/tflite_flutter, Accessed on 31st January 2023. 148

[267] Record: Flutter package. https://pub.dev/packages/record, Accessed on 31st January 2023. 148

[268] Eventify: Flutter package. https://pub.dev/packages/eventify, Accessed on 31st January 2023. 148

[269] matrix2d: Flutter package. https://pub.dev/packages/matrix2d, Accessed on 31st January 2023. 148

[270] Amna Anwar and Eiman Kanjo. Crime prevention on the edge: Designing a crime-prevention system by converging multimodal sensing with location-based data. In A Basiri, G Gartner, and H Huang, editors, *Proceedings of the 16th International Conference on Location Based Services (LBS 2021)*, pages 96–100, Glasgow, UK/online, 2021. TUWien. 158

[271] Amna Anwar, Eiman Kanjo, and Dario Ortega Anderez. Deepsafety:multi-level audio-text feature extraction and fusion approach for violence detection in conversations, 2022. 159

[272] Clive Norris and Michael McCahill. Cctv: Beyond penal modernism? *British Journal of Criminology*, 46(1):97–118, 2005. 219

[273] Joshua C Klontz and Anil K Jain. A case study of automated face recognition: The boston marathon bombings suspects. *Computer*, 46(11):91–94, 2013. 220

[274] Nurul Azma Abdullah, Md Jamri Saidi, Nurul Hidayah Ab Rahman, Chuah Chai Wen, and Isredza Rahmi A Hamid. Face recognition for criminal identification: An implementation of principal component analysis for face recognition. In *AIP Conference Proceedings*, volume 1891, page 020002. AIP Publishing, 2017. 220

[275] Imed Bouchrika, John N Carter, and Mark S Nixon. Towards automated visual surveillance using gait for identity recognition and tracking across multiple non-intersecting cameras. *Multimedia Tools and Applications*, 75(2):1201–1221, 2016. 220

[276] Wei Zeng and Cong Wang. View-invariant gait recognition via deterministic learning. *Neurocomputing*, 175:324–335, 2016. 220

[277] Diego Gragnaniello, Carlo Sansone, and Luisa Verdoliva. Iris liveness detection for mobile devices based on local descriptors. *Pattern Recognition Letters*, 57:81–87, 2015. 221

[278] Rutvik Kakadiya, Reuel Lemos, Sebin Mangalan, Meghna Pillai, and Sneha Nikam. Ai based automatic robbery/theft detection using smart surveillance in banks. In *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 201–204. IEEE, 2019. 221

[279] Medha Bhargava, Chia-Chih Chen, Michael S Ryoo, and Jake K Aggarwal. Detection of abandoned objects in crowded environments. In *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 271–276. IEEE, 2007. 221

[280] Teng Li, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan. Crowded scene analysis: A survey. *IEEE transactions on circuits and systems for video technology*, 25(3):367–386, 2014. 221

[281] Amie Schuck. American crime prevention: Trends and new frontiers. *Canadian Journal of Criminology and Criminal Justice*, 47(2):447–462, 2005. 221

[282] F. M. Javed Mehedi Shamrat, Sovon Chakraborty, Md. Shakil Moharram, Tonmoy Roy, Masudur Rahman, and Biraj Saha Aronya. A transfer learning approach for face recognition using average pooling and mobilenetv2. In Mukesh Saraswat, Harish Sharma, K. Balachandran, Joong Hoon Kim, and Jagdish Chand Bansal, editors, *Congress on Intelligent Systems*, pages 531–541, Singapore, 2022. Springer Nature Singapore. 221

[283] Shefiu Ganiyu, Olayemi Olaniyi, Olawale Surajudeen Adebayo, and Terfa Akpagher. Systematic review of facial recognition algorithms and approaches for crime investigations. 8:55–69, 2020. 221

[284] Samah A. F. Manssor and Shaoyuan Sun. Tirfacenet: Thermal ir facial recognition. In *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–7, 2019. 221

[285] Joe Purshouse and Liz Campbell. Privacy, crime control and police use of automated facial recognition technology. *Criminal Law Review*, 3:188–204, 2019. 222

[286] Marcus Smith and Seumas Miller. The ethical application of biometric facial recognition technology. *AI & SOCIETY*, 37(1):167–175, 2022. 222

[287] Daniel S Lawrence, Nancy G La Vigne, Margaret Goff, and Paige S Thompson. Lessons learned implementing gunshot detection technology: Results of a process evaluation in three major cities. *Justice Evaluation Journal*, 1(2):109–129, 2018. 223

[288] Bhagya Nathali Silva, Murad Khan, and Kijun Han. Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities. *Sustainable Cities and Society*, 38:697–713, 2018. 223

[289] Jerry H Ratcliffe, Matthew Lattanzio, George Kikuchi, and Kevin Thomas. A partially randomized field experiment on the effect of an acoustic gun-

shot detection system on police incident reports. *Journal of Experimental Criminology*, 15(1):67–76, 2019. 223

[290] Gerard Vivo-Delgado and Francisco J Castro-Toledo. Urban security and crime prevention in smart cities: a systematic review. 2020. 223

[291] Mun-su Park and Hwansoo Lee. Smart city crime prevention services: The incheon free economic zone case. *Sustainability*, 12(14):5658, 2020. 223, 227

[292] C Catlett, E Cesario, D Talia, and A Vinci. Spatio-temporal crime predictions in smart cities: a data-driven approach and experiments. pervasive mob. comput. 53, 62–74 (2019). 224

[293] Yu V Truntsevsky, II Lukiny, AV Sumachev, and AV Kopytova. A smart city is a safe city: the current status of street crime and its victim prevention using a digital application. In *MATEC Web of Conferences*, volume 170, page 01067. EDP Sciences, 2018. 224

[294] Hyunji Chung, Michaela Iorga, Jeffrey Voas, and Sangjin Lee. Alexa, can i trust you? *Computer*, 50(9):100–104, 2017. 225

[295] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, 2012. 225

[296] Jordan J Bird, Anikó Ekárt, and Diego R Faria. Learning from interaction: An intelligent networked-based human-bot and bot-bot chatbot system. In *UK Workshop on Computational Intelligence*, pages 179–190. Springer, 2018. 225

[297] Fred Richardson, Douglas Reynolds, and Najim Dehak. Deep neural network approaches to speaker and language recognition. *IEEE signal processing letters*, 22(10):1671–1675, 2015. 225

[298] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018. 225

[299] John M Blythe, Nissy Sombatruang, and Shane D Johnson. What security features and crime prevention advice is communicated in consumer iot device manuals and support pages? *Journal of Cybersecurity*, 5(1):tyz005, 2019. 225

[300] Chua Boon Liang, Mujahid Tabassum, Saad Bin Abul Kashem, Zulfiqar Zama, P Suresh, and U Saravanakumar. Smart home security system based on zigbee. In *Advances in Smart System Technologies*, pages 827–836. Springer, 2021. 225

[301] Huichen Lin and Neil W Bergmann. Iot privacy and security challenges for smart home environments. *Information*, 7(3):44, 2016. 225

[302] Fath U Min Ullah, Mohammad S Obaidat, Amin Ullah, Khan Muhammad, Mohammad Hijji, and Sung Wook Baik. A comprehensive review on vision-based violence detection in surveillance videos. *ACM Computing Surveys*, 55(10):1–44, 2023. 226

[303] Whistine Xiau Ting Chai and John Yu. Influence of social media on deviant acts: A closer examination of live-streamed crimes. In *Introduction To Cyber Forensic Psychology: Understanding The Mind Of The Cyber Deviant Perpetrators*, pages 23–44. 2021. 226

[304] Arijit Mukherjee, Arpan Pal, and Prateep Misra. Data analytics in ubiquitous sensor-based health information systems. In *2012 Sixth International Conference on Next Generation Mobile Applications, Services and Technologies*, pages 193–198, 2012. 227

[305] William D Burrell and Robert S Gable. From bf skinner to spiderman to martha stewart: The past, present and future of electronic monitoring of offenders. *Journal of Offender Rehabilitation*, 46(3-4):101–118, 2008. 228

[306] Eric Grommon, Jason Rydberg, and Jeremy G Carter. Does gps supervision of intimate partner violence defendants reduce pretrial misconduct? evidence from a quasi-experimental study. *Journal of Experimental Criminology*, 13(4):483–504, 2017. 228, 229

[307] Jo Brayford, Francis Cowe, and John Deering. *Sex Offenders: Punish, Help, Change Or Control?: Theory, Policy and Practice Explored*. Routledge, 2013. 228

[308] Mary A Finn and Suzanne Muirhead-Steves. The effectiveness of electronic monitoring with violent male parolees. *Justice Quarterly*, 19(2):293–312, 2002. 228

[309] Edna Erez and Peter R Ibarra. Making your home a shelter: Electronic monitoring and victim re-entry in domestic violence cases. *British journal of criminology*, 47(1):100–120, 2006. 228

[310] Marc Renzema and Evan Mayo-Wilson. Can electronic monitoring reduce crime for moderate to high-risk offenders? *Journal of Experimental Criminology*, 1(2):215–237, 2005. 229

[311] Rafael Di Tella and Ernesto Schargrodsky. Criminal recidivism after prison and electronic monitoring. *Journal of Political Economy*, 121(1):28–73, 2013. 229

[312] Stuart S. Yeh. Cost-benefit analysis of reducing crime through electronic monitoring of parolees and probationers. *Journal of Criminal Justice*, 38(5):1090–1096, 2010. 229

[313] Lawrence E Rothstein. Privacy or dignity: Electronic monitoring in the workplace. *NYL Sch. J. Int'l & Comp. L.*, 19:379, 1999. 229

[314] Guy Griffiths, Shane D Johnson, and Kevin Chetty. Uk-based terrorists' antecedent behavior: A spatial and temporal analysis. *Applied geography*, 86:274–282, 2017. 229

[315] Ajay Prasad, Kaushik Ghosh, and Sourabh Singh Verma. Crime patrolling assistance using passive monitoring: A proof of concept of a proactive wi-fi surveillance system. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 01–11, 2021. 230, 232

[316] Eiman Kanjo, Dario Ortega Anderez, Amna Anwar, Ahmad Al Shami, and James Williams. Crowdtracing: Overcrowding clustering and detection system for social distancing. In *2021 IEEE International Smart Cities Conference (ISC2)*, pages 1–7, 2021. 230, 231

[317] Julien Freudiger. How talkative is your mobile device? an experimental study of wi-fi probe requests. In *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pages 1–6, 2015. 230, 231

[318] Zhuliang Xu, Kumbesan Sandrasegaran, Xiaoying Kong, Xinning Zhu, B Hu, J Zhao, and C Lin. Pedestrain monitoring system using wi-fi technology and rssi based localization. *International Journal of Wireless & Mobile Networks*, 2013. 231

[319] Ajay Prasad, Kaushik Ghosh, and Sourabh Singh Verma. Crime patrolling assistance using passive monitoring: A proof of concept of a proactive wi-fi surveillance system. *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 00:01–11, 2021. 231

[320] Mohamad Huzaimy Jusoh, Muhammad Firdaus Bin Jamali, Ahmad Faizal bin Zainal Abidin, Ahmad Asari Sulaiman, and Mohamad Fahmi Hussin. Wi-fi and gsm based motion sensor for home security system apllication. *IOP Conference Series: Materials Science and Engineering*, 99(1):012010, 2015. 231

[321] Miguel Ribeiro, Nuno Nunes, Valentina Nisi, and Johannes Schöning. Passive wi-fi monitoring in the wild: a long-term study across multiple location typologies. *Personal and Ubiquitous Computing*, 26(3):505–519, 2022. 231

[322] Mohamad Huzaimy Jusoh, Muhammad Firdaus Bin Jamali, Ahmad Faizal bin Zainal Abidin, Ahmad Asari Sulaiman, and Mohamad Fahmi Hussin. Wi-fi and gsm based motion sensor for home security system apllication. *IOP Conference Series: Materials Science and Engineering*, 99(1):012010, 2015. 231

[323] Mathieu Cunche, Mohamed-Ali Kaafar, and Roksana Boreli. Linking wireless devices using information contained in wi-fi probe requests. *Pervasive and Mobile Computing*, 11:56–69, 2014. 231

[324] Dongsu Han, David G Andersen, Michael Kaminsky, Konstantina Papagiannaki, and Srinivasan Seshan. Access point localization using local signal strength gradient. In *International Conference on Passive and active network measurement*, pages 99–108. Springer, 2009. 231

[325] Brian Ferris Dirk Hähnel and Dieter Fox. Gaussian processes for signal strength-based location estimation. In *Proceeding of robotics: science and systems*. Citeseer, 2006. 231

[326] Sandra Wachter. Normative challenges of identification in the internet of things: Privacy, profiling, discrimination, and the gdpr. *Computer Law Security Review*, 34(3):436–449, 2018. 232

[327] C. Bisdikian. An overview of the bluetooth wireless technology. *IEEE Communications Magazine*, 39(12):86–94, 2001. 232

[328] Karl Benkic, Marko Malajner, P Planinsic, and Z Cucej. Using rssi value for distance estimation in wireless sensor networks based on zigbee. In *2008 15th International Conference on Systems, Signals and Image Processing*, pages 303–306. IEEE, 2008. 232

[329] Omotayo G Adewumi, Karim Djouani, and Anish M Kurien. Rssi based indoor and outdoor distance estimation for localization in wsn. In *2013 IEEE international conference on Industrial technology (ICIT)*, pages 1534–1539. IEEE, 2013. 232

[330] Abdalkarim Awad, Thorsten Frunzke, and Falko Dressler. Adaptive distance estimation and localization in wsn using rssi measures. In *10th Euromicro Conference on Digital System Design Architectures, Methods and Tools (DSD 2007)*, pages 471–478. IEEE, 2007. 232

[331] Tom Nicolai and Holger Kenn. About the relationship between people and discoverable bluetooth devices in urban environments. In *Proceedings of the 4th international conference on mobile technology, applications, and systems*

and the 1st international symposium on Computer human interaction in mobile technology, pages 72–78. ACM, 2007. 232

[332] Ryo Nishide and Hideyuki Takada. Detecting pedestrian flows on a mobile ad hoc network and issues with trends and feasible applications. *International Journal on Advances in Networks and Services*, 6(1&2), 2013. 232

[333] Rosalind W Picard. Automating the recognition of stress and emotion: From lab to real-world impact. *IEEE MultiMedia*, 23(3):3–7, 2016. 233

[334] Akane Sano and Rosalind W Picard. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 671–676. IEEE, 2013. 233

[335] Eiman Kanjo, Luluah Al-Husain, and Alan Chamberlain. Emotions in context: examining pervasive affective sensing systems, applications, and analyses. *Personal and Ubiquitous Computing*, 19(7):1197–1212, 2015. 233

[336] Yekta Said Can, Bert Arnrich, and Cem Ersoy. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics*, page 103139, 2019. 233

[337] Simon Ollander, Christelle Godin, Sylvie Charbonnier, and Aurélie Campagne. Feature and sensor selection for detection of driver stress. In *PhyCS*, pages 115–122, 2016. 234

[338] Yun Liu and Siqing Du. Psychological stress level detection based on electrodermal activity. *Behavioural brain research*, 341:50–53, 2018. 234

[339] Jennifer A Healey and Rosalind W Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6(2):156–166, 2005. 234

[340] Mark Colvin, Francis T Cullen, and Thomas Vander Ven. Coercion, social support, and crime: An emerging theoretical consensus. *Criminology*, 40(1):19–42, 2002. 234

[341] Sheilagh Hodgins. Mental disorder, intellectual deficiency, and crime: evidence from a birth cohort. *Archives of general psychiatry*, 49(6):476–483, 1992. 234

[342] Ann L Coker, Paige H Smith, Lesa Bethea, Melissa R King, and Robert E McKeown. Physical health consequences of physical and psychological intimate partner violence. *Archives of family medicine*, 9(5):451–457, 2000. 234

[343] Loring Jones, Margaret Hughes, and Ulrike Unterstaller. Post-traumatic stress disorder (ptsd) in victims of domestic violence: A review of the research. *Trauma, Violence, & Abuse*, 2(2):99–119, 2001. 234

[344] Donna Eileen Stewart and Simone Natalie Vigod. Mental health aspects of intimate partner violence. *Psychiatric Clinics*, 40(2):321–334, 2017. 234

Appendix A: Extended Literature Review

A.1 Audio-Visual Technologies

Audio and visual technologies, such as CCTV surveillance and facial recognition, have become increasingly popular and widely used for the prevention of crime in recent years. These technologies have the potential to provide real-time information and enable automatic detection and response to various threats to safety. The development of smart cities and smart home technologies has further expanded the capabilities of audio and visual technologies, enabling the integration and coordination of multiple sensors and systems. This section reviews the state-of-the-art in audio and visual technologies for crime prevention and discusses their potential and challenges in the context of CCTV surveillance, facial recognition, smart cities, and smart home technologies. In addition, multimedia (or multimodal) technologies are considered, which attempt to combine audio and video data for the purposes of crime prevention.

A.1.1 Video Technologies

Video technologies refers to technology that uses camera or video data to prevent crime. Crime prevention video technologies are wide ranging, with well-known systems such as CCTV being prevalent in modern society in both public and private scenarios. This section provides a review of this associated literature, considering how video technologies can be used for crime prevention in scenarios such as public safety, crowd control, and facial recognition. Therefore, video

technologies are used for both safety, crime prevention, and access control, highlighting their impact on modern society. For example, CCTV systems are the most popular choice of crime prevention technology around the world [46], which is due to the ease of use of setting up such systems, but also the effectiveness of the systems in capturing meaningful data.

A.1.1.1 CCTV Surveillance

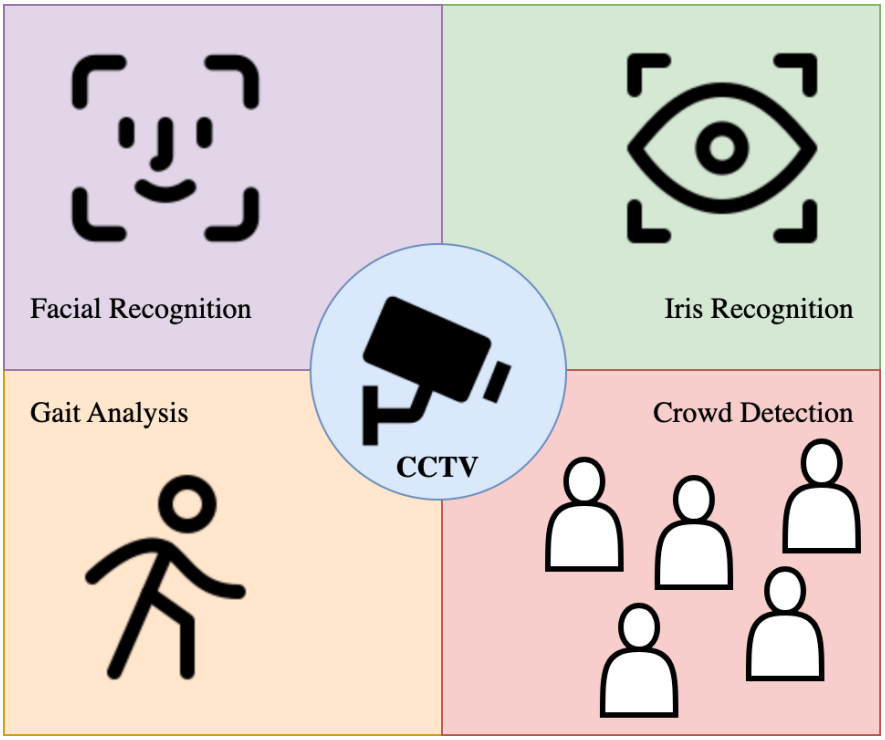

Figure 1: A diagram representing the different forms of CCTV application areas as themes that can be supported through digital video technologies for crime prevention.

In recent years, CCTV surveillance has emerged as the most popular choice of crime prevention technology worldwide [46]. It is a technology that is already readily available and installed in many public places such as airports, train station offices, school buildings, university buildings, and on the streets. In 2005 it was estimated that there are 4 million CCTV cameras installed in the United Kingdom, which is estimated in a ratio of 1 camera per 14 citizens [272]. Al-

though CCTV cameras have previously been known to be employed as a means of reporting and reducing crime by supporting police officers as a form of evidence in the court of justice or in the prosecution process. However, with increasing developments in computer vision and machine learning, it gives opportunities to use footage as a mechanism to prevent crime by investigating CCTV surveillance [46].

CCTV cameras are installed to provide safety and a mechanism to control crime rates. For example, offenders could be deterred by the fear of being on record while committing any crime. Another example of how CCTV cameras help law enforcement agencies intervene to reduce crime rates is by providing them with the opportunity to identify crimes while they are in progress or the ones that are about to happen. If these two instances fail, the recording from the cameras can be used as evidence for prosecution and to identify the people involved. Although all methods have been of use, they can still be ineffective if, for example, with the first example the offender does not perceive the CCTV as a risk to them and it will not stop them from committing the crime. The second mechanism requires staff monitoring to identify any suspicious behaviour in real-time to be able to take action; however, this can easily be missed if there are several camera feeds to monitor them. The last method requires the crime to occur in the camera view with clarity for the video footage to be of any use, and that could be a hindrance as that may require the camera to register the incident and have the camera follow it by tilting and moving. Hence, there are many factors that can emerge and, understandably, cause incidents to be missed due to human error. However, with constant development in the field, technology has been evolving in computer vision and machine learning, leading to continuous improvement in applications such as object and face recognition [273, 274] or gait analysis [275, 276], clearly providing opportunities for CCTV surveillance that could improve its effectiveness [46]. The uses of CCTV are identified in Figure 1.

With the constant change in technology, CCTV surveillance systems are being upgraded to incorporate the latest features of software technology [150], such as advanced feature extraction algorithms and machine learning classification algorithms. Therefore, it is now possible to record and extract personal biometrics and surveillance patterns that can be used a posteriori for the identification of people. CCTV technology has been around for a long time and therefore has been

investigated in different applications. These compromise of iris recognition-based security systems that deny access to buildings to unauthorised personnel [277], object and human posture detection for automatic robbery detection in banks [278], automatic detection of suspicious anomalies such as unattended bags in mass transit areas or crowded venues [279], and can also be a means of congestion analysis and crowd detection for safety [280].

A.1.1.2 Facial Recognition

Facial recognition technology has advanced in recent years and has been identified as a source of effective evidence-based policing since the early 2000s [281]. In more recent years, facial recognition has been integrated into mobile phones by leading smartphone providers, such as Apple and Samsung. This technology has also been proposed in several different systems that have been developed to unravel the faces of criminals through facial recognition technologies. Technologies used in facial recognition systems vary, with most systems implementing AI to detect, compare, and store faces in databases for future comparisons. A transfer learning approach has been described [282], which uses average pooling and MobileNetV2 to detect faces, which the authors describe as being for crime prevention. The authors preprocess the data on raw image data before testing the system. The model is successful when working with existing known face datasets using artificial light and achieves results of 98.89% and 99.01% on the training and test data. While successful, more work is needed in this regard, especially considering the lower quality, night environment, or obfuscation of an individuals face that could occur in outdoor environments or areas where crime prevention is needed more.

Previous work has presented a review of concepts related to face detection, computer vision, and facial recognition [283]. The analysis showed the benefits of using the CCTV camera feed for crime prevention and investigations. The categorisation highlights potential improvement from the model identified previously, as a description of each environment is provided to aid the design and development of facial recognition systems considering respective environmental factors. TIRFaceNet [284] investigate thermal Infrared facial recognition, which is demonstrated for security purposes in the detection of unauthorised individ-

uals at night. The authors propose a convolutional neural network (CNN) that uses preprocessed images of an individual's face, before the system extracts the features of the face (e.g., nose, eyes, mouth) with the system then training faces on the deep network. The features are then compared to a dataset, to detect unauthorised individuals. The results indicate positive results, with accuracies of 98.70% - 98.50% depending on the dataset used. Despite this, further work is needed to place such systems in a contextual environment, where uncertainty is common.

Previous research [71] has examined the use of digital police tools in predictive police, especially on the use of data-driven algorithms, facial recognition, and biometric identification technologies. This study examined the literature in several databases and selected 29 sources for analysis. Research has concluded that these technologies are increasingly being used for predictive police operations and more research is needed to understand the impact of these tools on police decision making. Several concerns have been raised about the privacy of individuals, with previous articles identifying inadequate protection afforded to privacy rights, and human rights of those subjected to facial recognition surveillance by the Police in England and Wales [285]. Therefore, it is necessary to keep the law up-to-date with technological innovation to ensure that facial recognition is used ethically and without infringing on the rights of individuals, and researchers recommend the ethical application of biometric facial recognition to support appropriate laws and regulations regarding the technology [286].

A.1.2 Audio Technologies

Audio-based technologies have been around for many years and have been used to anticipate different types of criminal action. A common example involves the interception of calls by the police force, wire tapping; or the collection mechanism as a form of evidence by recording devices [156]. The scope of audio-based technologies has experienced great expansion with the current advancement in audio processing and machine learning techniques, as several parameters and biometrics can be automatically computed from raw audio recordings. For example, the United States has recently evaluated the use of audio-based technologies for

the detection of gunshots [287]. Gunshot detection is a novel technology that employs a network of microphones that are mainly installed in areas with a high crime rate to minimise the chance of false gunshot detections, as it discriminates gunshots from other types of noise and computes the spatial coordinates of the location where the shot was fired from.

A.1.2.1 Smart Cities

Smart cities are on a rise and are becoming more and more common around the world to improve the quality of citizens' lives in large cities and make them more sustainable and streamline urban services using technology and data. Several major cities around the world, such as London, New York, Singapore, Barcelona, and Amsterdam, have adopted various degrees of Smart City Initiatives. These initiatives often focus on areas such as transportation, energy, and public safety. There are several challenges [288] that come with having smart cities, such as complexity and integration, sustainability, privacy and security, inclusion and equity, funding and governance, and cyber security.

For example, an article [289] discusses the use of gunshot detection systems such as ShotSpotter and its impact on the reporting of police incidents. The study found that the implementation of the ShotSpotter system increased the number of reports and arrests related to gunfire. Furthermore, the study found that the system had a positive effect on reporting gunshot incidents by police officers. The authors conclude that the implementation of gunshot detection systems can positively affect the accuracy and completeness of police reports.

A research project [290] explores the potential impact of smart city technology on police and law enforcement and how these technologies can be used to improve public safety and community relations. It also examines the potential challenges and ethical considerations associated with the use of these technologies in police. They argue that although smart city technology can improve public safety, it also raises important ethical and social concerns that must be addressed.

The study [291] examined the smart crime prevention services of the Incheon Free Economic Zone (IFEZ) in South Korea and identified ways to improve smart city security systems. The study collected IFEZ data in four functional areas and

10 scenarios between 2017 and 2018. The study found that in order to provide effective intelligent crime prevention services, the precision and coherence of the data must be verified, consistent processes must be established to link all crime prevention services, and experts from specialised institutions must be encouraged and ensured. The conclusions indicate that to develop intelligent city security services in an optimal manner, in-depth discussions about data collection and sharing are necessary.

A paper [292] focusses on urbanisation challenges, particularly in cities with high crime rates, and on how new technologies can help police access and analyse crime data to understand patterns and trends. The authors propose a predictive approach to detect high-risk areas of crime in urban areas and to forecast crime trends in each region with spatial analysis and automatic regression models. The result of this algorithm is a spatio-temporal crime forecast model that can estimate the number of crimes that may occur in a particular region. The methodology was tested using two real datasets from Chicago and New York City, and the results show that it is accurate in space and time crime forecasting.

The study [293] explored the problem of street crime in Russia and the potential use of modern digital technology to prevent it. The study found that street crime is a widespread problem in Russia and is on the rise, with a high proportion of mercenaries and violent crimes. It also found that street crime occurs mainly in empty urban spaces and is influenced by daytime, weekdays, and years. Furthermore, theft of mobile phones is a common problem and street crime is sudden and unpredictable. The study pointed out that the current way of preventing street crime was not effective and recommended that digital technology be used to improve it.

A.1.2.2 Smart Home Technologies

Computational detection of violent language is of interest due to its potential use in smart home devices; across a variety of domestic locations, for example kitchens, lounges and other living spaces. These living spaces provide a variety of scenarios in which smart home devices can be used, for example, to order ingredients, change lights, and record family moments.

Likewise, with the presence of Chatbots or Intelligent Virtual Assistants (IVAs) such as Google Assistant and Amazon Alexa, are becoming increasingly popular around the world [294]. Such IVAs incorporate speech recognition capabilities that allow users to ask questions and make requests to different interfaces. In addition to speech recognition, it is now possible to count the number of speakers in a conversation by speaker diarisation [295], to infer the sentiment (mood) of individuals by analysing their voice [296], or to recognise a speaker by their voice with considerably low error rates [297, 298]. These advances clearly indicate the success of their respective areas and provide the opportunity for them to be implemented in audio-based crime prevention tools.

However, with such technologies also comes a significant security risk. For example, [299] discusses the security risks of the Internet of Things (IoT) and how it can threaten user security, privacy, and safety. The study showed that many IoT devices on the market lacked security features and consumers did not always use the available security features. They analysed 270 consumer IoT devices' user manuals and support pages from 220 different manufacturers, and it was found that manufacturers provide limited information about the security features of their devices, making it difficult for consumers to make informed decisions about the safety of devices before purchasing.

Smart home technologies can offer many things, for example, a project that [300] aims to create low-cost smart home security systems using Zigbee technology to protect Sarawak residents from flooding, smoke and invasion. The system includes hubs, battery sensor nodes, and Android apps. Sensor nodes send alerts to the hub and the Android app when detection of intrusion, flooding, or smoke. The system also captures intruder photos, has live mobile monitoring, and is controlled by the Internet.

Another paper [259] discusses the security concerns of existing home automation systems and their inability to prevent sophisticated intruders. Highlighting the need for advanced technology to identify and protect homes from skilled intruders. It also emphasises the importance of security in the development and implementation of home automation systems to provide residents with a sense of security.

This article [301] examines the IoT and the privacy and security needs of

different applications. It identifies that the needs of critical engineering infrastructures and sensitive commercial operations differ from those of the smart home environment in the house. It studied existing solutions to improve IoT security and identified future key requirements for reliable smart home systems. Gateway architectures are considered to be best suited for resource-limited devices and high system availability. The project suggested two key technologies to ensure safe system operation: automatic system configuration and automatic update of system software and firmware.

A.1.3 Multimedia Technologies

Multimedia technologies, such as live streaming and video sharing platforms, have become increasingly popular and widely used in recent years. These technologies have the potential to be valuable tools for crime prevention, allowing real-time monitoring and detection of criminal activities. Recent studies have explored the use of multimedia technologies for crime prevention. For example, a study used machine learning to automatically detect and classify violent scenes from videos [302].

This paper [117] presents a modular approach for law enforcement authorities to effectively manage and process large amounts of heterogeneous data generated by digital technologies to support criminal investigations. The proposed platform will use new technologies and efficient components capable of transferring and disseminating multimedia information and provide a single secure point for analysing and multidimensional visualisation of criminal information. The increased volume, mode, and frequency of the data generated by tools and individuals makes it critical that law enforcement agencies use this information effectively.

A chapter called “Influence of Social Media on Deviant Acts: A Closer Examination of Live-Streamed Crimes” in the “Introduction to Cyber Forensic Psychology” book discusses the impact of social media on illegal and criminal acts, especially live streaming technologies [303].It also explores how live streaming may contribute to facilitating or facilitating criminal behaviour and the possible impact of this phenomenon on crime prevention and law enforcement. The chap-

ter also highlights the perception and experiences of live broadcasting crime and its potential impact on social norms and values.

In general, these studies show that multimedia technologies are effective in crime prevention and can provide real-time information and automatically detect criminal activity. However, more research is needed to assess the effectiveness and limitations of these technologies and address potential ethical and privacy concerns.

A.2 Ubiquitous Sensing

Ubiquitous sensing refers to the use of a wide range of sensors and other technologies to collect data from the physical environment in real time and to enable various applications and services. Existing research [291] suggests that ubiquitous sensing can be a valuable tool to improve safety in cities, providing real-time information and allowing automatic detection and response to various threats to safety. However, the authors also recognise that there are challenges and limitations to the use of this technology for crime prevention, such as privacy concerns, cost, and reliability.

Previous work [258] discusses the various types of sensors and technologies that are used in ubiquitous sensing, such as wireless sensors, cameras, and microphones. The authors also discuss various applications of ubiquitous sensing, such as health care, transportation, and environmental monitoring. Similar work [304] provides a comprehensive overview of the state-of-the-art in ubiquitous sensing technology for healthcare applications. The authors discuss the various types of sensors and technologies used in this context, as well as their potential applications and benefits. In general, the paper suggests that ubiquitous sensing has the potential to enable a wide range of applications and services for health care, but that there are significant challenges that need to be addressed to realise its full potential.

EM and GPS are technologies that allow for the tracking and location of individuals or objects in real-time [53]. These technologies have been widely used for various applications, including criminal justice, transportation, and logistics. In recent years, the development of short-range sensing technologies, such as Wi-

Fi and Bluetooth Low Energy (BLE), has further expanded the capabilities of EM and GPS, enabling new applications and opportunities. This section covers the basics of long-distance sensing, and discusses the potential of short-range sensing technologies in this context.

A.2.1 Long-Distance Sensing

Long-distance sensing refers to sensors, whose main purpose is to communicate information over long distances consistently but through a wireless network. Due to the nature of some offences, it is necessary for technologies to effectively and continuously map the signal to ensure that victims are safe and that any parole or similar rules are being followed. This has traditionally been performed using EM, which is a method of supervision through constant location monitoring [305]. Through more modern methods such as GPS [306] or connecting to mobile networks, it is possible to augment the data collected by an EM device to be more precise and mapped for effective surveillance.

A.2.1.1 Electronic Monitoring

EM began in the early 1980s in the US and spread rapidly after positive initial claims in reducing control deficits in community supervision [305]. The first EM systems employed radio frequency identification (RFID) technology. RFID technology is based on a tag with a unique identifier that sends data to an electronic reader through wireless radio frequency waves, allowing identification and tracking of it. RFID technology was first employed to confine offenders (or pre-trial defendants) to a particular location (usually their home) [307]. The work in [308] showed that placement of sex offenders in EM programmes reduced the probability of their return to prison and postponed their return to prison. The use of RFID technology was then extended to the protection of victims of domestic violence. Victims were provided with receivers so that they were alerted when an offender was present within a pre-established control perimeter, which was normally set to be around the victim's home [309] (see also Fig. 2).

Concerns about the effectiveness and privacy of EM have also been considered; for example, a study on the effectiveness of EM in reducing crime was not

supported by existing data [310], the authors identified the need for future studies to investigate and draw conclusions on the effectiveness of technology in reducing crime before recommendations can be made regarding the use of technologies. A similar conclusion has also been found in a study on EM after prison in Argentina, in a comparison between two judges who allocate EM differently [311], with the authors identifying a large and negative effect on criminal recidivism compared to the case of prison. Despite this, other work has considered the cost-benefit analysis of using EM technologies, where data from a national survey were used to estimate the costs and benefits on EM technologies [312]. The authors found that EM technologies would avert an estimated 781,383 crimes every year, which would provide a social value of the reduction in crime of $481.1 billion, highlighting the potential of technology to be cost effective for wide scale deployments. There is only a limited selection of literature on EM effects on the privacy of those it is provided to, however, work has been conducted considering the effect of EM in the workplace, where it is considered a potential privacy risk when implemented for non-crime related populations [313].

A.2.1.2 Global Positioning System

The second generation of EM technologies incorporated the use of GPS. GPS is a satellite-based global navigation system that provides geolocation through the use of a network of satellites orbiting the Earth at an altitude of approximately 20,200 km. To estimate the geo-location, a GPS receiver intercepts the signals of at least three network satellites at regular intervals of time. A posteriori, based on the time it takes to receive each of the satellite signals, the geo-location of the GPS receiver is calculated via trilateration. The use of GPS technology as a crime prevention tool has gained increasing attention since the late 1990s [306]. The ability to customise exclusion zones and provide instant alerts when they are violated has extended the use of electronic monitoring to sex offenders and post-work release offenders [51]. A pictorial example of how GPS technology is used in this context can be seen in Fig. 2. More applications, such as tracking terrorist suspects to gain insight into their spatial and temporal behaviour, have recently been proposed [314].

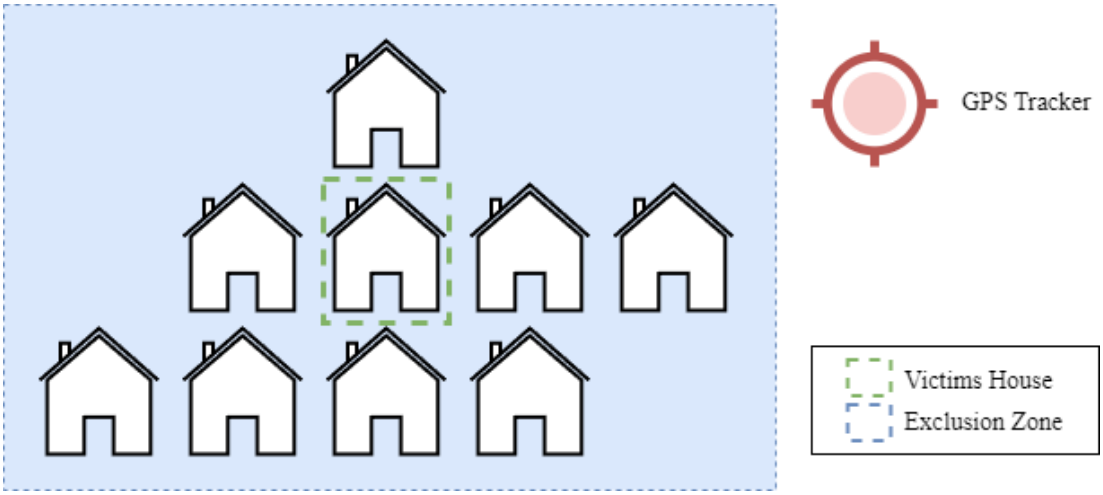

Figure 2: Diagram presenting how GPS geo-location can be applied to electronic monitoring in the context of urban or suburban neighbourhoods. The illustration presents an exclusion zone around the victims house, with the GPS tracker being outside of the exclusion zone.

A.2.2 Short-Range Sensing

Short-range wireless communication transceivers, such as Bluetooth and WI-FI are widely incorporated into many portable and mobile devices, including laptops, mobile phones, and smart watches. During the manufacturing process, a wireless module is assigned a unique identification (ID) in the form of a 48-bit Medium Access Control (MAC) address [315]. This address is then used to identify and authenticate a device when communicating with other wireless devices. Short-range sensing makes use of the MAC address or Bluetooth identifiers to count the number of devices within range of the sensors, this enables approximations as to specific devices or the number of individuals in a crowd [316]. The dimensions of the crowd can then be used for several applications, including for the purposes of crime prevention. Despite this, concerns relating to the privacy of such implementations are common, with manufacturers implementing MAC address randomisation to counter tracking methods using probe requests [317].

A.2.2.1 Wi-Fi

Wireless sensing is an emerging area of information tracking in the population; wireless sensing is the process by which information is collected about the public through wireless signals (e.g., [318–321]). Data capture techniques that are possible often focus on monitoring crowds or individuals based on data collection requirements. For example, Wi-Fi sensing research will often use aggregation of individual device clusters, making it useful for crowd sensing, such as in the case of crowd detection for Covid-19 overcrowding prevention [316]. Although Wi-Fi sensing can be used to detect wireless packets in transit by individual devices, there are limited examples of this; most Wi-Fi sensing work has a focus on large groups, and Bluetooth sensing on individual devices. There is a small focus on crime prevention, such as detecting emerging problematic crowds or using it as a method of detecting intrusions of specific Wi-Fi devices on a local network [322].

There are two main approaches to using Wi-Fi for sensing; these can be identified as the use of individual Wi-Fi probe requests for node-node communication and Wi-Fi access point scanning for identifying the number of devices connected to a hub:

- Wi-Fi probe requests refers to the use of Wi-Fi packet sniffing and intercepting to count and collect data from IEEE802.11 Wi-Fi probe requests [317], which can be used to collect large amounts of data according to the packets shared by mobile devices. This data, while useful in context, cannot capture rich details about individuals due to MAC address randomisation. Despite this, the ability to link data from similar devices is possible [323], including using methods such as Wi-Fi fingerprinting or crowd analysis.

- Wi-Fi access point scanning is the scanning of public wireless access points and monitoring changes in signal strength to gather estimation insights [324, 325] on the size of a crowd. This method uses access points that already exist in most locations, which can then be linked to a wider network area.

Although only minor examples of the use of Wi-Fi sensing technologies in crime prevention are described in the existing literature, the potential use of

such technologies in practise is widespread, with commercial entities using this data to analyse the urban environment and co-locate potential shopfronts. Questions about the effectiveness of these approaches in direct identification and privacy concerns related to General Data Protection Regulation (GDPR) commitments [326] may be the reason why crime prevention professionals consider alternative and more interoperable approaches. There are still opportunities to deploy existing resources to emerging situations [315], however, there are questions about the usefulness of this approach over the more widespread and reliable CCTV networks available in most urban areas.

A.2.2.2 Bluetooth

Bluetooth devices can interact with other nearby Bluetooth devices within their signal range ($10m$ to $100m$, depending on the radio transceiver) by sending and receiving radio waves within a band of 79 different frequencies centred at 2.45 GHz. Using such radio waves, along with the identification capabilities provided by the unique MAC address assigned, a Bluetooth device can continuously monitor other Bluetooth devices nearby (within its signal range) and also identify the type of device associated with such MAC address (for example, whether it is a smart phone or a laptop) [327]. The Received Signal Strength Indicator (RSSI) provides an estimated measure of the power present in a received radio signal. As shown in previous research [328–330], RSSI can be used to estimate the approximate distance that a Bluetooth receiver is from a Bluetooth emitter.

Using the above characteristics of this technology, various researchers have employed Bluetooth technology to estimate the social context surrounding a person [160] (see Fig. 3) or to estimate pedestrian flows in specific locations [331,332]. The authors of [161] discuss how such monitoring capabilities could be used to prevent or investigate infant and elementary school kidnapping. They note that kidnappings tend to take place when children are alone and that a system could be used to continuously log the Bluetooth devices near a child and, when none are detected, an accelerometer is used to monitor their activity. Ultimately, they suggest, a mobile application (where the Bluetooth logs can be visualised) could be provided to the child's parents so that they can monitor their children's social

context throughout the day.

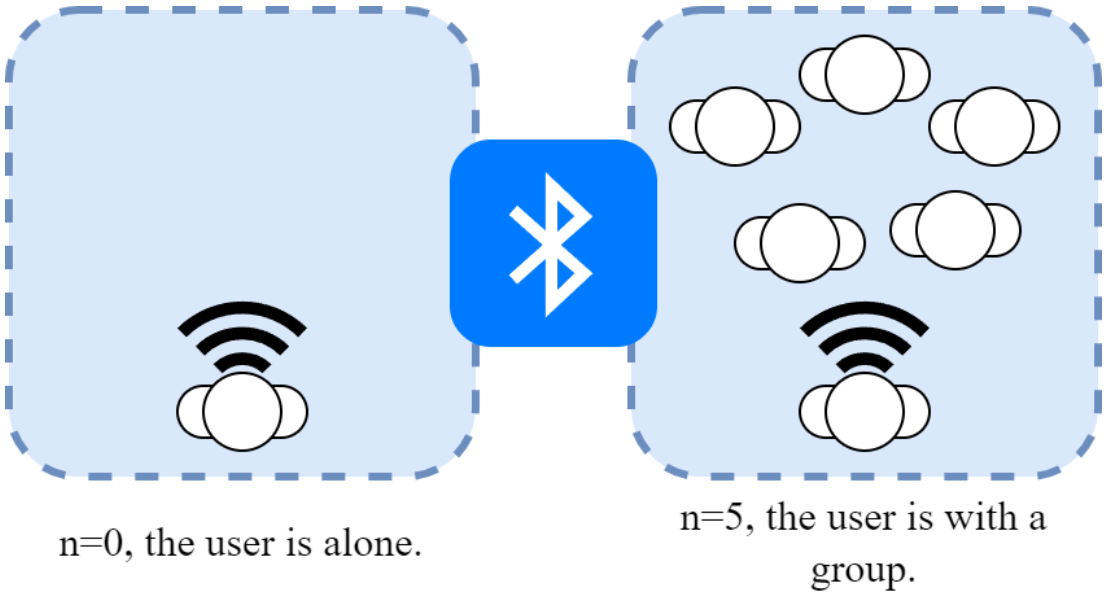

Figure 3: A illustration of social context monitoring via Bluetooth device scanning. The examples presented in the diagram display a n=0 situation where no other devices are scanned, and a n=5 situation where multiple other devices are scanned.

A.2.3 Affective Computing

Affective computing is a field of study based on designing and developing technology that can recognise, interpret and respond to human emotions. In the context of crime prevention, affective computing could potentially be used to develop systems that can identify individuals who are at risk of committing a crime, based on their emotional state and behaviour. Affective computing research has become aware that there is a relationship between physical health and emotional state of an individual [333] and given this, the field has gained a significant amount of attention in the past few years [334–336].

Affective computing, or emotional intelligence, is the study and development of systems for the recognition, processing, and interpretation of human affects. Typically, this is performed with the use of wearable devices or smart textiles by which various physiological signals related to stress levels are measured, processed

and interpreted. Electrodermal activity and heart rate variability are two main examples of physiological signals used in affective computing. Although affective computing is still in its infancy, several studies have shown that such signals can be translated into relevant features, which ultimately lead to estimates of human stress levels. For example, research studies in [337,338] have used the MIT Stress Recognition in Automobile Drivers Database [339] to classify between three different stress levels (low, medium and high) in three different driving scenarios.

Affective computing has not been shown to be used in the context of crime prevention when conducting a literature review; however, anticipating high levels of stress could help prevent crime. The literature in the field suggests that both mental disorder and violence can be caused by the stress an individual experiences [340]. Mental disorders have been shown to increase the chances of committing a crime. For example, in the study conducted by [341] using a random Swedish birth cohort, it was found that men with major mental disorders were more than ten times more likely than men without mental disorders to have a criminal history and four times more likely to be registered for a violent offence. Similar research [342–344] suggests that on average victims of domestic violence have poorer mental health, which can lead to other problems, such as depression or anxiety.