National College of Ireland

A Classification Approach to Identifying Female Victims of Intimate Partner Violence in Europe and the US

MSc Research Project
Data Analytics

Megan Farrelly
Student ID: x19144440

School of Computing
National College of Ireland

Supervisor: Dr. Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Megan Farrelly
Student ID:	x19144440
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr. Vladimir Milosavljevic
Submission Due Date:	19th September 2022
Project Title:	A Classification Approach to Identifying Female Victims of Intimate Partner Violence in Europe and the US
Word Count:	11,175
Page Count:	27

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	17th September 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Classification Approach to Identifying Female Victims of Intimate Partner Violence in Europe and the US

Megan Farrelly
x19144440

Abstract

Violence against women is a global public health issue. However, it can be difficult to recognise if a woman is suffering abuse. Intimate partner violence is a subset of this abuse and few studies focus on identifying abuse caused by a partner or ex-partner using classification techniques. The studies that have been completed tend to apply the same models but report different performance metrics. Little work has been done to determine how to improve these models within this domain. To address these knowledge gaps, this study applied nine classification models, including ensemble, boosted, stacked and deep learning techniques, to determine which model was most appropriate to identify women suffering intimate partner violence. It was found that Random Forest returned the highest accuracy and AUC. XGBoost, Support Vector Machine and a stacked classifier also returned favourable metrics, while the deep learning techniques tended to perform poorly. It was found that reducing the number of features input into a Random Forest model reduced the average accuracy returned but maintained the error of the model. By reducing the number of data points input, accuracy was maintained but error increased. The findings suggest that geography plays a role in the rate of violence suffered, along with the number of people living in the same household. These are novel findings which may aid future classification studies and guide them when resources such as data availability are limited. Overall, it is hoped that this research will aid stakeholders such as healthcare professionals or women's charities, to improve risk assessments or to identify women at risk of abuse using an appropriate model identified by this study. The recommendations made by this study could also alleviate the issues faced by women suffering violence by facilitating better government and business decisions at a community level.

1 Introduction

Violence against women is considered a major public health issue, affecting women globally each day. According to the World Health Organisation (WHO), 30% of women worldwide experience violence, most commonly in the form of intimate partner violence (IPV)[1]. Violence against women is defined as any act of gender based violence that results in physical, sexual or psychological harm of women. IPV is considered a subset of

[1]https://www.who.int/news-room/fact-sheets/detail/violence-against-women

this, where an intimate partner or ex-partner is responsible for the harm[2]. This abuse can have a profound effect. Zara and Gino (2018) found, when analysing violent killings of women in Italy over a 45 year period, that more than half of the victims were in an intimate partnership with the perpetrator. It has also been found that children of IPV victims are significantly more likely to die within five years of being born (Rawlings and Siddique, 2020). To make matters worse, the COVID-19 pandemic has exacerbated the issue, increasing the rate of IPV and violence suffered by women (van Gelder *et al.*, 2020; Xue *et al.*, 2020). A solution to this problem must be found.

There have been attempts within the domain of predictive analytics to solve the issue of violence against women. Rodríguez *et al.* (2021) reported that there was a dramatic increase in the number of published studies from 2018 to 2020 where computer based techniques were applied to address violence against women. Anderson *et al.* (2021) found through a systematic review that identifying victims of IPV through mobile based devices is low cost and enables support of victims who are part of a minority. Despite the explosion of studies in recent years, it is believed that gaps within the literature exist. By addressing these gaps, the benefits cited by Anderson *et al.* (2021) could be provided.

This study will address the following research question: *"To what extent can machine and deep learning techniques identify female victims of intimate partner violence in Europe and the United States, and can an appropriate predictive model and the factors influencing it be identified for use as a risk assessment tool for support providers?"*. A number of objectives have been set to answer the research question. These are:

- To conduct a critical and thorough review of the pertinent literature to identify the knowledge gaps and inform the scope of the project.
- To apply nine different classification models to two datasets describing IPV against women in Europe and the US.
- To compare the performance of the models applied by reviewing multiple evaluation metrics.
- To compare the performance of the models to performance reported in the published literature using corresponding evaluation metrics.
- To determine the factors that may influence a woman's risk of IPV.
- To identify the most appropriate model for potential deployment as an IPV risk assessment tool for community use.

It is hoped by answering this research question, that novel and topical research that addresses identified knowledge gaps and drives future work within this domain is complete. The major contribution of this research will be a comprehensive comparison of multiple classification techniques modelled on data describing IPV, including novel applications, which future studies identifying IPV can refer to. Additionally, by identifying the most appropriate model and the factors which influence IPV, a risk assessment tool to detect female victims of IPV in the community may be identified.

[2]https://www.ohchr.org/en/instruments-mechanisms/instruments/declaration-elimination-violence-against-women

2 Related Work

Machine and deep learning studies within the domain of IPV and crimes committed against women from the years 2016 to 2022 were reviewed to identify knowledge gaps and inform the scope of the research. This informed domain knowledge and confirmed that answering the research question would result in a novel contribution to the research area.

2.1 Social Media Data

Within the current literature, online detection studies account for over 50% of computer-based violence against women studies (Rodríguez *et al.*, 2021). Mostly, these studies apply techniques to data scraped from social media. Subramani *et al.* (2018) applied multiple deep learning techniques, including a Convolutional Neural Network (CNN), a Recurrent Neural Network, a Gated Recurrent Unit (GRU) and a Long-Short Term Memory (LSTM) to Facebook data describing critical instances of domestic violence. They found that GRU performed best, with an accuracy of 94% and precision and recall rates of 95%. In a similar study, Rodríguez-Sánchez *et al.* (2020) applied deep learning techniques Bidirectional Encoder Representations from Transformers (BERT) and LSTM, along with machine learning techniques Linear Regression, Random Forest and Support Vector Machine to Spanish tweets to detect sexist posts. This study reported that BERT returned an accuracy of 74%. García-Díaz *et al.* (2021) applied machine learning techniques including Sequential Minimal Optimisation (SMO) to Spanish tweets also describing misogynistic content, with SMO returning an accuracy of 85%. This is interesting as deep learning techniques are often considered to be more robust than machine learning techniques yet these studies suggest that machine learning techniques fare better when detecting female victims of abuse online.

A deep learning multilayer perceptron was applied to Mexican tweets describing gender based violence (Castorena *et al.*, 2021). They achieved Area Under the Receiver Operating Characteristic Curve (AUC) of 80%, suggesting good balance between specificity and sensitivity of correctly classified gender based violence tweets (Castorena *et al.*, 2021). These studies show that social media data can be classified using both machine and deep learning techniques, with the potential to reduce violence against women online. Castorena *et al.* (2021) note that hate speech on social media platforms is often quickly flagged and removed, yet the same is not done for posts describing abuse against women. But in order to identify victims using the techniques described above, victims must publicly post an account of the abuse they are suffering, which may be unlikely to happen. Rodríguez *et al.* (2021) found that of the online detection studies, over half focus on detecting misogyny directed at women online, while only 14% focus on detecting reports of abuse. Hence, there is a need for this research, where female victims will be identified within the community without having to expressly declare their abuse.

2.2 Factors influencing IPV

In order to understand the nuances in the results returned by predictive models, it is important to fully understand the factors that influence a woman's risk of becoming a victim of IPV. Many studies have been completed, typically using regression, to identify these factors. McFarlane *et al.* (2016) applied Logistic Regression to predict the likelihood

of a victim returning to a shelter to develop a triage tool so that those most at risk may be provided better support. They found that age, if a child witnessed the abuse, the relative risk of murder and the tangible support available were the most important factors influencing the model. The regression model returned an AUC of 90%, suggesting the model could distinguish between classes well. However, they report an adjusted R^2 value of 27%, suggesting that the model does not fit the data well (McFarlane *et al.*, 2016). Amusa *et al.* (2020) also applied Logistic Regression, along with Random Forest, Decision Tree and Gradient Boosting to data on married women to identify their risk of suffering IPV and the factors that influence that risk. They found that fear of their partner was the most important predictor of IPV. Raj *et al.* (2021) applied Logistic Regression to national survey data to predict sexual violence outside of partnerships. They found that exposure to past violence and lack of sexual and reproductive health knowledge were predictors of sexual violence. These results suggest that sexual violence outside of marriage may be reduced by educating women on sexual health. A more recent study further builds on this, using the same techniques and survey data but to predict IPV, finding suffering emotional violence and being separated from their partner increased a woman's likelihood of seeking help (Dehingia *et al.*, 2022). Through regression analysis, Borraz and Munyo (2020) found that increasing women's social welfare payments, and therefore improving their economic independence, significantly reduced domestic violence rates. In contrast, Alesina *et al.* (2021) found through regression that economically independent women in Africa were more likely to suffer domestic violence. This suggests that factors that influence the rate of IPV vary depending on nuances specific to geographic regions.

Research has also been conducted to determine the factors that influence the justification of IPV. Sáez *et al.* (2020) found through ANOVA that alcohol consumption by a victim of IPV increased both self-blame and external blame attribution. Hossain *et al.* (2022) determined through regression that burning meals, refusing sex and going out without informing their partner were prominent justifications for IPV.

It seems many different aspects of a woman's life and her relationship with her partner can affect her risk of becoming a victim of IPV, as well as the proportion of blame attributed to her. This study aims to identify the factors that attribute to the models making a positive classification. It is wondered if similar factors will be returned to those identified in the literature.

2.3 Machine Learning Studies

Machine learning techniques have been previously applied to different data sources from within the community to predict the risk of violence against women in different situations. These data include demographic data, as well as geographical and image data, often collected through national surveys. Rodríguez-Rodríguez *et al.* (2020) applied Logistic Regression, Random Forest and k-Nearest Neighbour following feature selection to data describing gender based violence, including the number of murders and calls to victim support helplines, to predict the number of gender based violence complaints brought to court. It was found when modelling for the entirety of Spain, Random Forest returned the lowest RMSE value of 17% (Rodríguez-Rodríguez *et al.*, 2020). Chen *et al.* (2020) also applied Logistic Regression and Random Forest to radiology reports describing women's injuries to predict IPV, with Random Forest returning an AUC of 85%. The data modelled in these studies is in contrast to the data proposed by this study, predicting complaints brought to court or physical injuries caused by IPV, with the intention of

tackling IPV through better policy planning. These studies showcase the wide range of use cases for predicting IPV using classification techniques.

As proposed by this study, attempts have been made to predict IPV with data collected within the community. As mentioned, Amusa *et al.* (2020) applied multiple models to identify married women's risk of suffering IPV, reporting that Random Forest outperformed other models with an AUC of 76%. However, they also report that this model returned an accuracy of 53%, while Decision Tree returned an accuracy of 65%. Hossain *et al.* (2021) found that Random Forest, Logistic Regression and Naïve Bayes (NB) algorithms predicted family violence including IPV in Bangladeshi homes during the pandemic with accuracies of 77%, 69%, and 62% respectively. Raj *et al.* (2021) found that applying Logistic Regression to predict sexual violence outside of partnerships returned an AUC of up to 86%. Dehingia *et al.* (2022) achieved an AUC of up to 83% using the same techniques and survey data as Raj *et al.* (2021), but to predict IPV. Reza *et al.* (2021) modelled six machine learning techniques including XGBoost, Ada Boost, Decision Tree and Random Forest to data describing area, year and women and child oppression to develop a support system that forecasts crime based on input data. XGBoost performed best with an R^2 of 99%. With the exception of Amusa *et al.* (2020), Hossain *et al.* (2021) Reza *et al.* (2021), limited studies compare several techniques and fewer report multiple metrics. It is noted that there is an inconsistency between the metrics reported in the above studies, making it difficult to compare these studies to each other. It will be interesting to identify the differences and similarities between the results of this research and these studies.

2.4 Deep Learning Studies

Few deep learning techniques have been utilised to predict violence against women. This may be due to the fact that the most appropriate algorithm for predicting this specific crime has yet to be identified. Zhang *et al.* (2020) investigated the most appropriate model for predicting crime, particularly public property crime in China. The authors applied CNN and LSTM, along with k-Nearest Neighbours, Random Forest, Naïve Bayes and Support Vector Machine, and concluded that LSTM outperformed the other models, likely due to its ability to extract time-series trends (Zhang *et al.*, 2020). The authors further built on this study in 2022, in order to identify an interpretable model with a high degree of accuracy in predicting crime (Zhang *et al.*, 2022). The authors note that typically regression models have been used due to the ability to determine the contribution of each predictor variable, but machine learning techniques tend to be more accurate and therefore appropriate (Zhang *et al.*, 2022). Often machine learning algorithms are considered black boxes as it is not clear how they arrive at the end prediction. Zhang *et al.* (2022) found XGBoost returned an accuracy of 89% and AUC of 59%, and an age of 25 to 44 to be the greatest contributing factor in committing public theft. XGBoost is a boosted machine learning algorithm, so despite their success with deep learning in 2020, Zhang *et al.* (2022) decided an interpretable machine learning technique was more appropriate for predicting crime. These are interesting studies that may explain why few deep learning techniques have been used to predict crime, including gender-based crimes, as they can be difficult to interpret despite their performance power.

In a similar study, He and Zheng (2021) applied deep learning techniques to predict crime hot spots, using Generative Adversarial Network to output images of predicted hotspots on maps. Baek *et al.* (2021) applied two deep learning models including CNN to

a text-based summary of crimes received by police, including crimes typically committed against women such as rape, to predict crime type and risk. CNN returned an accuracy and F1 score of 91% and 84% respectively. The authors note that CNN has the ability to return an accuracy 7% greater than Support Vector Machine (Baek *et al.*, 2021). Safat *et al.* (2021) applied LSTM for time series analysis to predict crime type and rates, which returned RMSE values as low as 9%. While deep learning techniques have yet to be applied with the intention of predicting crime against women, particularly IPV, they have been successfully applied to general crime data. It is wondered if similar performance will be achieved in this research. However, it is again noted that the metrics reported between studies predicting crime are inconsistent, making direct comparisons of performance between studies difficult.

2.5 Identified Gaps

From the above review, it has been found that multiple gaps within the current body of knowledge exist. First, the majority of studies focus on identifying abuse against women online, particularly using text-based social media data, while few focus on detecting abuse within the community (Rodríguez *et al.*, 2021). Second, most studies use regression techniques to identify the factors that influence the rate of IPV. It is wondered if different results could be returned with more robust feature importance techniques. Third, there is a leaning towards the application of the same few techniques within the machine learning studies, such as Random Forest and Logistic Regression. Fourth, there is a lack of consistency in the evaluation metrics reported between studies. Fifth, few studies compare multiple techniques with the sole objective of identifying IPV. Sixth, no studies have been published that focus on identifying IPV over large geographical areas. Amusa *et al.* (2020) and Reza *et al.* (2021) compare multiple techniques in the identification of IPV within South-Africa and Bangladesh respectively. It is noted that there appears to be no published work on identifying European or US victims of IPV, while studies with an interest in geographic regions tend to focus on Asian countries. This research may identify an appropriate technique for identifying European or US victims of IPV that differs from the published literature. Seventh, it is noted that these studies do not examine and compare the relative error returned by each model. While the reported results imply successful classifications, the variance in results returned may fluctuate greatly, thus making it difficult to identify the most appropriate model. Finally, no studies appear in the published literature where deep learning techniques are applied to predict IPV.

By addressing these gaps, it is hoped that the research question will be answered and an accurate tool in identifying women currently or at risk of suffering IPV can be identified, similar to the tool described in the work by McFarlane *et al.* (2016). This tool could potentially provide more-tailored support to victims or aid decision making by stakeholders.

A summary of the primary studies that have informed the scope of this research is presented in Table 1.

3 Methodology

A comprehensive research methodology was established and followed to answer the research question. This methodology was based on the Cross-Industry Standard Process

Table 1: The primary studies predicting IPV against women. The techniques applied in these studies are are Random Forest (RF), Gradient Boosting (GB), Decision Tree (DT), Logistic Regression (LR), Naïve Bayes (NB), XGBoost (XGB) and AdaBoost (AB).

Authors	**Techniques**	**Data**	**Metrics**	**Results***
Amusa *et al.* (2020)	RF, GB, DT, LR	Residency, ethnicity, education	AUC	76%
Hossain *et al.* (2021)	RF, LR, NB	Income, education, age, residency	Accuracy	77%
Reza *et al.* (2021)	XGB, AB, DT, RF	Area, year	R^2	99%
Dehingia *et al.* (2022)	LR	Education, age, residency, religion	AUC	83%

*Result of best performing model reported

for Data Mining (CRISP-DM) process, a structured approach for data mining research that is well established within the field. This process was chosen as the basis of the methodology as CRISP-DM is an iterative process that enables research through steps that interact with and assist each other. By tailoring this process to this research, a methodology was constructed that facilitated the timely completion of the analysis, returned the major deliverables and allowed for the research question to be answered. The methodology can be viewed in Figure 1.

3.1 Domain Understanding and Data Exploration

Following acquisition of the data and identification of the research question, the first step in the research methodology was to obtain an understanding of predictive analytics in

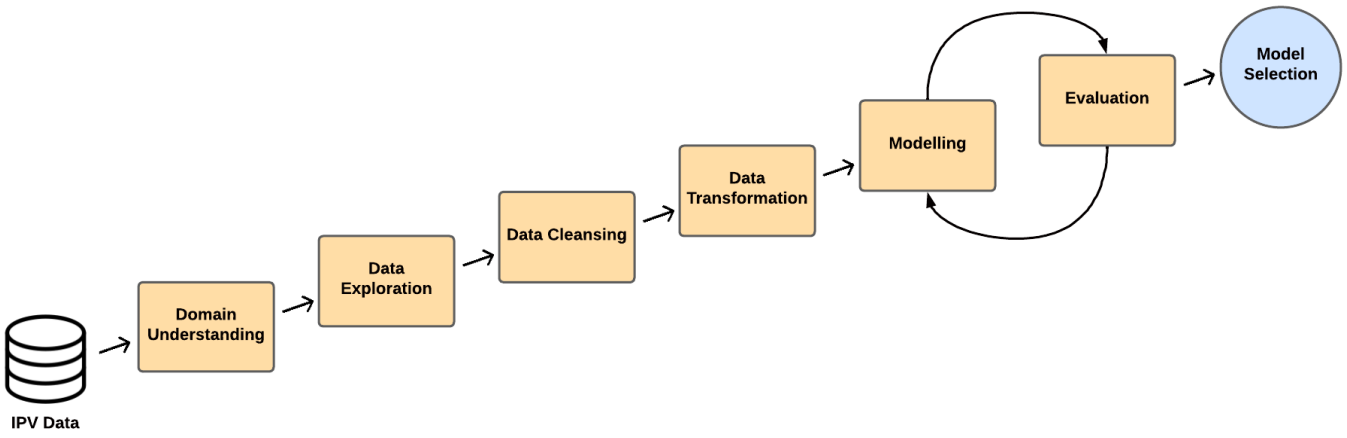

Figure 1: The methodology followed to complete the research.

the domain of violence against women. This was achieved through the comprehensive literature review which identified the knowledge gaps and steered the direction of the research. By understanding the domain and the preceding research, a basis was formed for understanding the data.

Two datasets were obtained to complete the research. First, a private secondary dataset describing violence against women in Europe was obtained. This dataset was collected by the European Union Agency for Fundamental Rights (FRA) and is published as the Violence Against Women Survey, 2012 dataset[3]. This dataset is not publicly available, therefore an application was made to the UK Data Service, the data controller of these data, and permission for use in this research was granted by the owners. This dataset contains data from a survey conducted with 42,000 women across 28 European countries in 2012. The type of violence, if any, that each woman suffered, along with demographic data such as age, education and income status, were collected. This dataset contains 42,002 data points and 3,417 variables.

The second dataset obtained was public data describing crime in the USA. This secondary dataset was collected by the United States Bureau of Justice Statistics office and is published as the National Crime Victimisation Survey (NCVS) by the Inter-university Consortium for Political and Social Research[4]. This survey focuses on the victims of crime in the United States between 1992 and 2020. The dataset includes data describing crimes against women, as well as the victims' demographic data such as age, race and income status. No studies have been published which use the NCVS dataset with a focus on violence against women. This dataset contains 297,399 data points and 1,017 variables.

Following procurement of the data, exploration was carried by reading the accompanying documentation and data dictionaries. Variables of interest were identified and noted. Exploration identified 48 variables of interest in the FRA dataset and 30 in the NCVS dataset. The data were imported to the research environment for further exploration. The number of null values in each dataset were determined. As well, the variables of interest were further examined and redundant variables were identified. The data types of each variable were determined. The number of values within each variable were explored. Obvious redundant values were noted. As data were obtained via survey, values such as "Not applicable" or "Don't know" were common throughout both datasets if an interviewee could not answer a question. The number of these values were identified and noted for each variable. Histograms and correlation plots were generated to visualise the distribution of the data and the relationships between each variable. Skew and kurtosis were determined for numeric variables. Using the FRA data, the rate of IPV in each European country was explored. Following exploration, cleansing began.

3.2 Data Cleansing and Transformation

Null and redundant values were dropped from the data. Variables that were found to contain mostly nulls or were highly correlated with other variables were removed. Manipulation was required to create the target variable for each dataset. The target variable was defined as whether the interviewee suffered IPV or not. As only women were interviewed in the FRA study, where an interviewee reported that they had either been physically or sexually abused, or both, by a partner or ex-partner was coded as IPV, and all other

[3]https://bit.ly/3sVmbeg
[4]https://bit.ly/3t5B8u7

instances were coded as not IPV. For the NCVS data, where both male and females who suffered a crime were interviewed, data points where the gender was recorded as male were removed. Then the target variable was created by coding where a female suffered abuse committed by a partner or ex-partner as IPV, while all other instances were coded as not IPV. This abuse included rape or attempted rape, physical or sexual assault or verbal threats of abuse. Finally, the values in each remaining variable were renamed so that feature importance could be analysed later. IPV was found to make up 12.54% and 21.22% of the target variable of the FRA and NCVS datasets respectively.

Following cleansing, the data were transformed for machine and deep learning. As the majority of variables were categorical, encoding was required. First, the data were split into train and test sets. Transformations were only fit to the training sets to avoid data leakage. Non-ordinal categorical variables were one-hot encoded and ordinal categorical variables such as education level were ordinal encoded. Numerical variables were normalised by scaling the values between zero and one. To solve class imbalance within the target variables, Synthetic Minority Oversampling Technique (SMOTE) Tomek was applied to generate synthetic data points of the minority class (Batista *et al.*, 2003). SMOTE Tomek was chosen as it addresses both over and undersampling by reducing the number of overlapping data points. The resulting transformed training sets contained 22,974 data points and 134 features and 24,720 and 16 features for the FRA and NCVS data respectively. Following cleansing and transformation, application of machine and deep learning models to both datasets could begin.

3.3 Modelling and Evaluation

Nine models were selected to predict IPV, including a bagging method, a stacked method, two boosted methods and two deep learning methods. These models were:

1. Naïve Bayes: Chosen as it is simple to implement and understand, and to compare performance to Hossain *et al.* (2021).

2. Decision Tree: Chosen as it also simple to implement and easily understood, and to compare performance to Amusa *et al.* (2020).

3. Support Vector Machine: Chosen as it works well on data with non-linear relationships, and to determine if this method performs as well in detecting IPV as it does in detecting online abuse against women (Rodríguez-Sánchez *et al.*, 2020; García-Díaz *et al.*, 2021).

4. Random Forest: Chosen as it is a powerful ensemble method that can also provide insight into feature importance. This model was found to be the most popular model in predicting abuse of any kind against women, while Amusa *et al.* (2020) determined it to be the most appropriate model for predicting IPV. It was chosen so comparisons to these studies could be made.

5. XGBoost: Chosen as it is a powerful boosted technique that can provide insight into feature importance and return high accuracies (Chen and Guestrin, 2016). Additionally chosen so performance could be compared to Reza *et al.* (2021).

6. Light Gradient Boosting Machine: Chosen as this technique has yet to be applied to data describing abuse against women and to compare performance against XGBoost to determine the effectiveness of boosted methods on these data (Ke *et al.*, 2017).

7. Stacking Classifier: Chosen as this technique has yet to be applied to data describing abuse against women, and to determine the effectiveness of a stacked method on these data in comparison to more popular techniques.

8. Multilayer Perceptron: Chosen as no published studies focus on applying deep learning techniques to data describing IPV, and to determine the effectiveness of this method on these data in comparison to more popular techniques. This feed-forward, fully connected network can be a powerful classifier when an appropriate activation function is used.

9. TabNet: Chosen as deep learning techniques typically do not perform well on tabular data (Ryan, 2021). TabNet is a deep learning architecture designed specifically to address this issue and can provide insight into feature importance (Arik and Pfister, 2021). The architecture consists of fully connected layers, batch normalisation and gated linear units. This model was also chosen to compare performance against multilayer perceptron, to determine the effectiveness of deep learning methods on these data.

The metrics chosen for the study were accuracy, AUC and F1 score, to provide an overall indicator of model performance through sensitivity, specificity, precision and recall, as well as accuracy of the model. Multiple metrics were chosen to facilitate in-depth performance analysis and to allow comparisons to previous studies that report differing metrics. The test set for each dataset was bootstrapped by resampling the data 200 times with replacement. The average accuracy, AUC and F1 score was determined, along with the 2.5^{th} and 97.5^{th} percentile to determine the error for each model by creating a bootstrapped or non-parametric confidence interval at the 95% level (Pace, 2012). These metrics were recorded for each model and were evaluated. The parameters of each model were tuned for both datasets, and modelling and evaluation were repeated until optimal results were achieved and the most appropriate model for both datasets could be chosen. Feature importance analysis was conducted for appropriate models, and error analysis was conducted by altering the dimensionality of the data to determine the drivers of variance within the models.

4 Design Specification

Practical research was carried out according to the architecture described in the design specification in Figure 2. All research was carried out on Mac OS Monterey on a MacBook Air with a 1.1 GHz Quad-Core Intel Core i5 processor and 8 GB of RAM. The FRA data were acquired in TAB format while the NCVS data were acquired in TSV format. Both datasets were initially imported to Jupyter Notebook but the environment was changed to Google Colab as it was later found that a GPU accelerator was required. Both datasets were converted to dataframes upon importation and were cleansed and transformed in the Python 3 programming language. The pandas and numpy packages were commonly used for these tasks. Following this, exploratory analysis, model application and error analysis were conducted. Commonly used packages in the application stage were scikit-learn, tensorflow, keras and pytorch. Once results were recorded for each model, they were analysed and compared. Visualisation was accomplished using the plotly package and Tableau 2022.1. Through metric comparison and appropriate graphing, the most appropriate model for identifying IPV could be selected for potential community use.

This three tiered design specification allowed for the model and the influencing features to be identified, along with the factors driving the error.

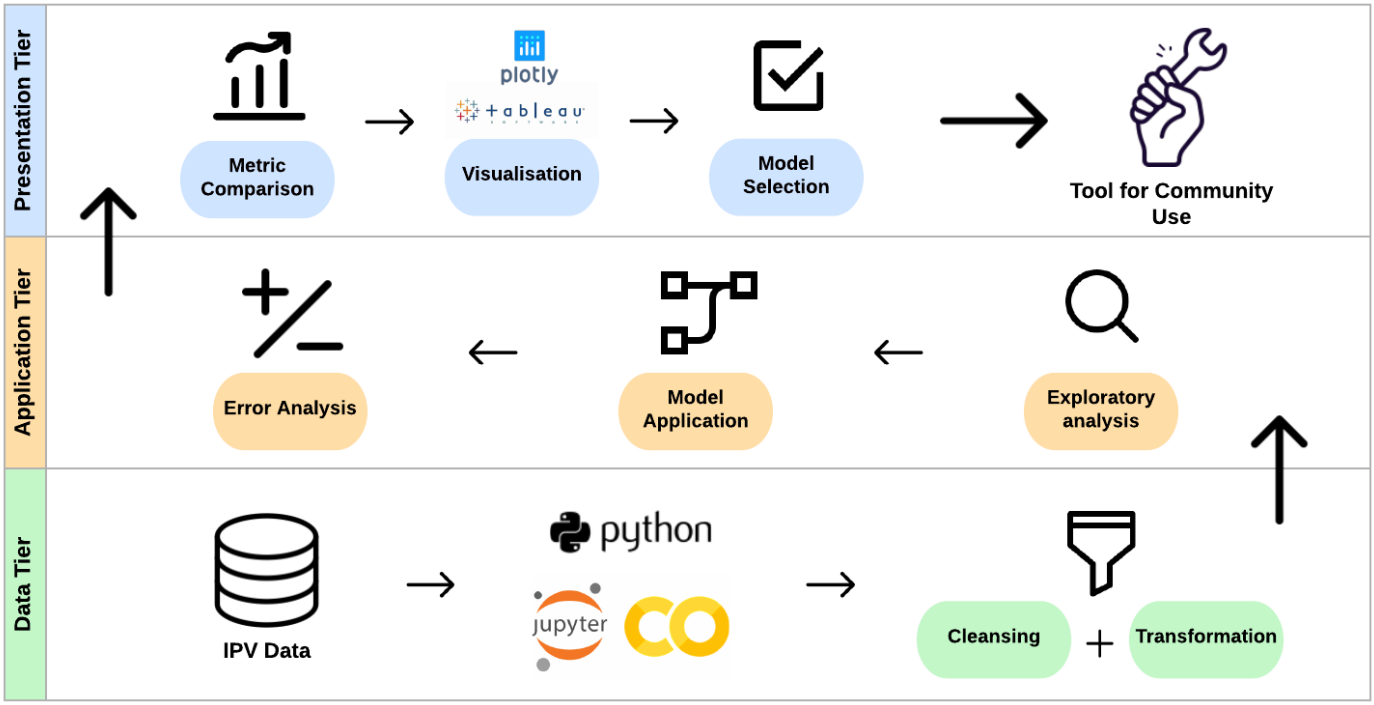

Figure 2: The architecture that supported the research.

5 Implementation

By following the methodology and design specification discussed, the intended solution to the research question could be implemented. The required packages were imported and the data were uploaded to Google Colab to be explored, cleansed, transformed and modelled. To enable efficient programming, functions were created for each repeated task. First, the European FRA data were explored for outliers, nulls and other redundant data. Following experimentation with imputation, the decision was made to remove null values, along with other values that provided little information, such as "Not applicable" or "No answer." Features that contained too many nulls or redundant values were excluded from the analysis to maintain performance of the models. Descriptive statistics for each feature were generated. The number of complete cases was calculated until an acceptable amount for modelling was achieved through cleansing. The target variable was created by combining two columns describing physical and sexual abuse using the numpy package. The percentage of each class within the target variable was calculated to determine whether there was class imbalance. The values of each variable were renamed from their coded value to a descriptive value so that feature importance could be easily completed later on. The resulting output of cleansing was a dataframe containing 16,418 rows and 40 variables to be transformed prior to modelling. Additionally, the rate of IPV in each European country was explored through stacked barcharts created with plotly.

Following cleansing, data were subset into training and test sets prior to transformation, at an 80% and 20% split respectively. These subsets were passed to a function that transformed the data to be fit for model application. This function one hot encoded non-ordinal categorical variables, ordinal encoded ordinal categorical variables, normalised numeric variables, and label encoded the target variable. This function also applied SMOTE Tomek to the training sets to solve class imbalance. The function was designed to only fit transformation techniques to the training sets, and once fit, to then apply the techniques to both sets. This was to prevent data leakage. The outputted FRA training set contained 22,974 training points and 134 features. Model application could then begin.

The optimal inputs of each model were tested thoroughly. Throughout this optimisation, accuracies of the training and test sets were calculated to prevent overfitting. Categorical Naïve Bayes from scikit-learn was applied to the transformed FRA training data, as was Decision Tree, Support Vector Machine with a radial basis function kernel, and Random Forest. Boosted methods XGBoost and Light Gradient Boosting Machine from the XGBoost and LightGBM libraries respectively were applied. A stacked model compiled of three initial estimators, Naïve Bayes, Support Vector Machine and Random Forest, and one final estimator, XGBoost, was constructed and applied. Two deep learning methods were applied. A multilayer perceptron consisting of one input layer, two hidden layers and one output layer was constructed, with one input node for each feature in the input layer. Relu and sigmoid activation functions were used in the different layers, and an adam optimiser was applied throughout. Training occurred with batch size 10 and 20 epochs. TabNet with an adam optimiser was applied to the training data, with a batch size of 32 and 10 epochs. The test set was used to calculate multiple metrics for each model. Accuracy, AUC, F1 score and a confusion matrix were outputted for each. The test data was bootstrapped with replacement so that each model was tested 200 times and the averages and variance of these metrics were outputted. The importance of each feature was calculated when Random Forest, XGBoost and TabNet were applied. The top 15 features for these models were identified and plotted in a barchart or boxplot. Graphs describing loss and accuracy per epoch on the training set were output by the deep learning models.

The same process was followed for the NCVS data. The target variable was created by combining values where the crime was committed by a spouse, partner or ex spouse or partner, and the crime was abuse, including rape, assault or verbal threats of abuse. Data were explored and cleansed to a dataframe containing 20,402 rows and 12 variables to be transformed prior to modelling. This dataframe was subsetted into a training and test set of 80% and 20% respectively. The training set was then transformed using the same function that transformed the FRA data, resulting in a training set of 24,720 training points and 16 features. The same nine models were applied following identification of the optimal parameters. Multinomial Naïve Bayes was applied instead of Categorical, and a polynomial kernel returned optimal performance for Support Vector Machine. A four-layer multilayer perceptron with input nodes for each feature, relu and sigmoid activation functions and adam optimiser was constructed and trained on batch size 32 and 10 epochs. TabNet was trained on batch size 20 and 10 epochs. The test set was bootstrapped and the average accuracy, AUC and F1 were returned for each model. Following exploratory analysis, modelling and feature importance analysis, error analysis was complete on the FRA data. A Random Forest model that reproduced the same results through setting a seed was established. Features in the training set were reduced by approximately a

quarter, a half and three quarters by removing features at random. The training points were reduced by taking random samples of a quarter, a half and three quarters of the size of the training set. Training points were also artificially doubled, tripled and quadrupled through random sampling with replacement. These subsets were transformed and passed to the model and accuracy and variance were recorded to determine whether feature or training point size were influencing the error returned by the models. This was the process that was followed to implement the methodology.

6 Results

The results obtained by following the above methodology are presented below. Exploratory analysis, modelling, feature analysis and error analysis all returned interesting results that were crucial to answering the research question.

6.1 Exploratory Analysis

The rate of IPV across all European countries is presented in a stacked barchart in Figure 3. The barchart is ordered according to rate of IPV, with countries to the left reporting higher rates than countries to the right. Hungary was found to have the highest rate of IPV, with over 20% of Hungarian participants reporting having suffered IPV. Ireland was found to have the lowest rate of IPV, with over 6% of Irish participants reporting having suffered IPV.

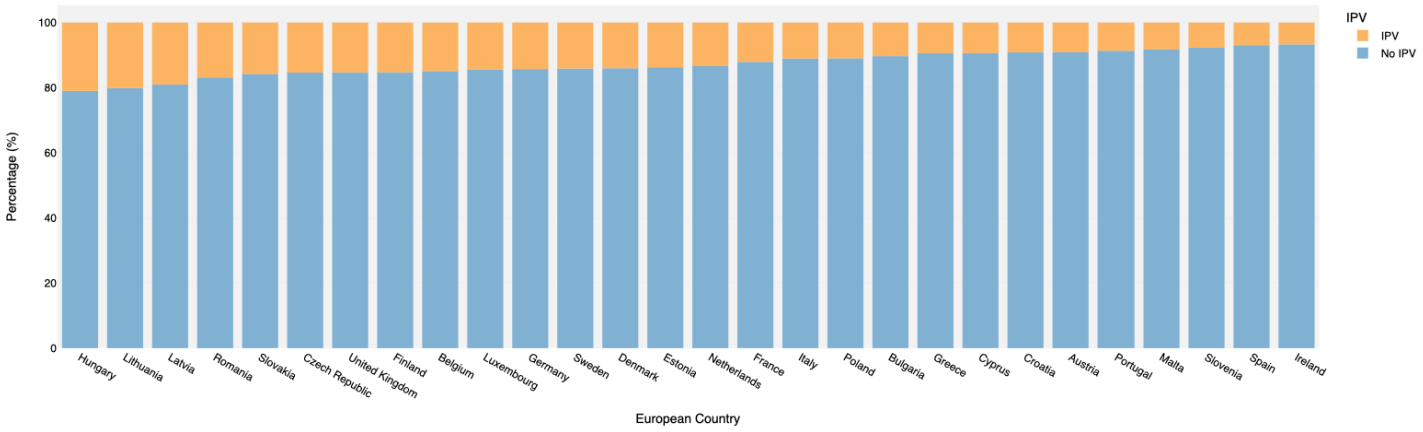

Figure 3: The percentage rates of IPV per country.

6.2 Model Results

The results of modelling are presented in Table 2. Accuracy, AUC and F1 are reported for all nine models on both datasets. Random Forest was found to return the highest accuracy and F1 score when modelled on the FRA data. A high F1 score suggests that both precision and recall are high, and that this model can correctly identify all positive cases. Support Vector Machine returned the highest AUC when modelled on the FRA data, meaning this model could distinguish between classes the best, although the value is moderate suggesting that either sensitivity or specificity is low. XGBoost returned the

same F1 score as Random Forest, along with a similar accuracy score. Light Gradient Boosting Machine returned the same accuracy as XGBoost, with a higher AUC than XGBoost and Random Forest, but a lower F1 score. Multilayer perceptron returned the lowest accuracy and F1 score for the FRA data.

The stacked model returned the highest accuracy and F1 score when modelled on the NCVS data. Random Forest returned the highest AUC when modelled on the NCVS data. Naïve Bayes modelled on the NCVS data performed the worst out of all models across the two datasets, returning the lowest accuracy and F1 score overall, as well as the lowest AUC for the NCVS data, despite returning a decent accuracy of over 80% when modelled on the FRA data. Overall, the models tended to return higher accuracies and F1 scores when modelled on the FRA data, but higher AUC with less variation when modelled on the NCVS data.

Table 2: The metrics of each model and the bootstrapped confidence interval at the 95% level for both datasets. The techniques applied in these studies are are Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), XGBoost (XGB), Light Gradient Boosting Machine (LGBM), a stacked classifier, multilayer perceptron (MLP) and TabNet (TN). The highest returned metric for each model is underlined.

Model	**FRA Data**			**NCVS Data**		
	Accuracy (%)	**AUC (%)**	**F1 (%)**	**Accuracy (%)**	**AUC (%)**	**F1 (%)**
NB	83.5 ± 1.2	61.3 ± 2.1	90.6 ± 0.8	56.8 ± 1.6	60.8 ± 1.6	66.3 ± 1.6
DT	81.3 ± 1.2	58.3 ± 2.0	89.3 ± 0.7	75.5 ± 1.3	66.8 ± 1.8	84.1 ± 0.9
SVM	77.6 ± 1.5	68.8 ± 0.2	86.3 ± 1.1	61.8 ± 1.6	63.2 ± 1.9	71.5 ± 1.5
RF	88.0 ± 1.1	53.5 ± 1.2	93.5 ± 0.6	79.8 ± 1.1	68.2 ± 1.6	87.3 ± 0.8
XGB	87.9 ± 1.1	54.6 ± 1.4	93.5 ± 0.6	74.4 ± 1.3	62.8 ± 1.6	83.6 ± 1.0
LGBM	87.9 ± 1.1	55.9 ± 1.8	93.4 ± 0.6	78.1 ± 1.4	62.0 ± 1.5	86.6 ± 0.9
Stacked	87.4 ± 1.1	53.6 ± 1.5	93.2 ± 0.6	80.1 ± 1.2	66.7 ± 1.6	87.7 ± 0.8
MLP	71.5 ± 1.5	67.3 ± 2.5	81.7 ± 1.2	63.5 ± 1.3	63.8 ± 1.5	73.2 ± 1.1
TN	78.9 ± 1.4	63.2 ± 2.8	87.5 ± 0.9	70.1 ± 1.3	64.3 ± 1.7	79.7 ± 1.1

6.3 Feature Analysis

Feature importance was analysed for Random Forest, XGBoost and TabNet for both datasets. As Random Forest was found to be the top performing model on the FRA data, and outperforming XGBoost and TabNet on the NCVS data, the top 15 features as identified by Random Forest permutation feature importance for both datasets is presented in Figures 4 and 5. Features such as health, length of the partnership and how often the abuser was intoxicated with alcohol were identified as the top features in the FRA data when predicting IPV (Figure 4). Meanwhile, the number of members in the household and victim's age were identified as having the largest effect in the prediction of IPV with the NCVS data (Figure 5). XGBoost, which returned similar metrics, found if the victim lived in a town to have the largest effect when predicting IPV with FRA

data. Similar to the top two features identified in Figure 5, XGBoost found the number of household members under and over the age of 12 were the top two features in the prediction of IPV with NCVS data. TabNet found if the victim was married at least once, if they worked a desk based job, if they were currently married and their age to be the top features in the prediction of IPV between the two datasets.

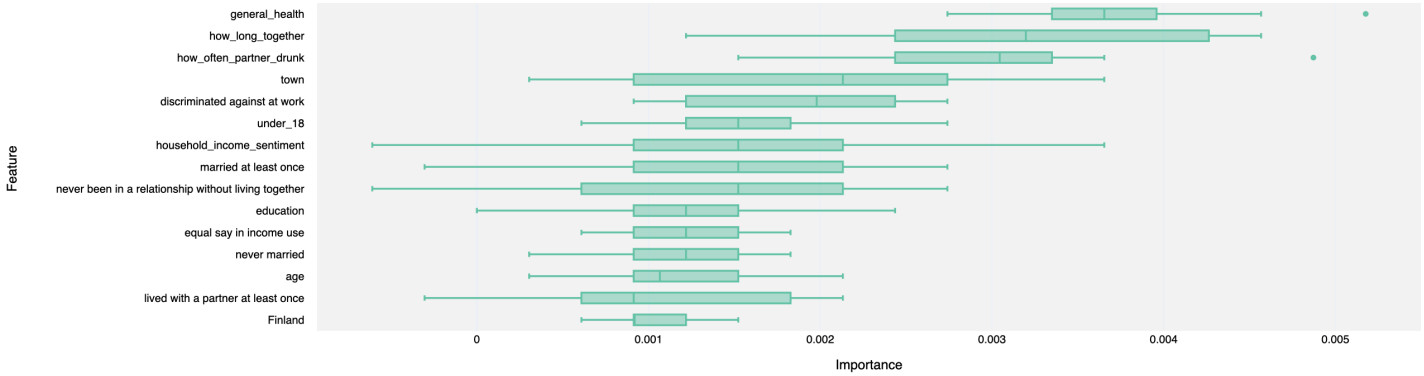

Figure 4: The top 15 features as identified by RF on FRA data.

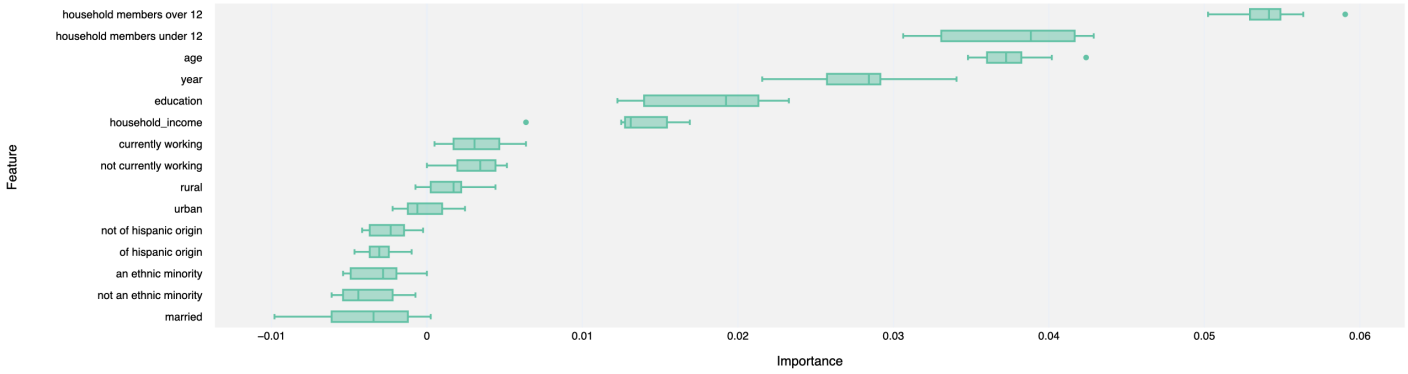

Figure 5: The top 15 features as identified by RF on NCVS data.

6.4 Error Analysis

Table 3 presents the results of Random Forest modelled on the FRA data when the 134 features and 22,974 training points were reduced by a quarter, half and three quarters, and the training points were doubled, tripled and quadrupled. Variance in the average accuracy returned remained relatively stable as features were reduced, but average accuracy decreased. When training points were reduced, variance increased while the average accuracy remained stable. When training points were increased, accuracy increased while variance in returned error was reduced.

7 Discussion

Overall, the results of this research suggest that an appropriate model for classifying whether a woman will suffer IPV can be chosen with confidence. As well, the features

Table 3: Mean accuracy and variance (%) at the 95% confidence level for Random Forest on FRA data when features and training points are reduced and training points are increased. The number of training points are rounded to the nearest thousand.

	Features	**34**	**67**	**100**	**134**
Reduction		79.9 ± 1.3	84.0 ± 1.1	86.3 ± 1.1	87.8 ± 1.2
	Training Points (thousands)	**6**	**11**	**17**	**23**
		87.5 ± 2.2	87.3 ± 1.4	87.1 ± 1.2	87.8 ± 1.2
Increase	**Training Points (thousands)**	**23**	**46**	**69**	**92**
		87.8 ± 1.2	97.3 ± 0.4	98.9 ± 0.2	99.4 ± 0.1

that should be collected to improve model performance and the amount of data required to return minimal error can be identified.

7.1 Exploratory Findings

Exploratory analysis found that Hungary, Lithuania and Latvia had the highest rates of reported IPV, while Ireland, Spain and Slovenia had the lowest. This is not so surprising, considering that Hungary, Lithuania and Latvia are some of the few European countries that have yet to ratify the laws proposed by the Istanbul Convention, a human rights treaty that the European Union signed in 2011[5]. Conversely, Spain and Slovenia were some of the first countries to ratify this treaty, so the low rate of IPV in these countries may be explained by their willingness to legislate against such abuse[6]. Ireland is an outlier, as despite reporting the lowest rate of IPV, they are one of the latest countries to ratify the Istanbul Convention. The strong opposition in these Central European and Baltic countries to ratify laws against violence against women has been attributed to conservatism, traditional ideologies and right-wing politics[7]. However, the acceptance of the Istanbul Treaty cannot fully explain the figures displayed in Figure 3, as this survey was conducted in 2012, years before it was ratified by any country. Weighted factors also cannot explain these results, as there were approximately equal amounts of data per country following cleansing, and Spain and Ireland have larger populations than Lithuania and Latvia[8]. It is thought that this difference is most likely due to differing culture and politics. Lithuania and Latvia, being neighbouring countries, likely share

[5]https://www.coe.int/en/web/conventions/full-list?module=signatures-by-treaty&treatynum=210
[6]See Footnote 5.
[7]https://emerging-europe.com/news/emerging-europe-domestic-violence-istanbul-convention/
[8]https://www.statista.com/statistics/685846/population-of-selected-european-countries/

some cultural views which then affects government policies implemented. These states of the former USSR may maintain a traditional Soviet mindset, comparable to Russia where abuse against women is not often acknowledged or addressed[9].

The Scandinavian countries Finland, Sweden and Denmark reported a higher level of IPV than was expected, appearing on the higher end of the graph (Figure 3). It was later found that these countries often report high levels of IPV despite having some of the highest rates of gender equality in the world. This is known as the Nordic Paradox and is yet to be explained (Gracia and Merlo, 2016). It must be considered that women from some countries may be more willing to share that they have been abused than others, and that the rate of IPV in Ireland may be much higher than reported. As these results are based on self reported abuse, their validity not only relies on the participants' willingness to honestly answer whether they have been abused or not, but also their understanding that they have been abused. It could be possible that women of some countries are more aware of what constitutes abuse by a partner than others. Further data collection and analysis following a mass awareness program would be required to determine the true rate of IPV across Europe. However, Gracia and Merlo (2016) note that disclosure rates do not fully explain the differences seen in reported IPV between countries. The results presented in Figure 3 suggest that geography, most likely driven by culture, education and government policies, play a role in the rate of IPV suffered by women.

7.2 Model Selection

From the results presented in Table 2, it can be seen that Random Forest outperformed other models, followed by XGBoost, Light Gradient Boosting Machine, the stacked classifier and Support Vector Machine. Interestingly, multilayer perceptron returned the lowest accuracy and F1 score when modelled on the FRA data. This may provide some insight as to why deep learning techniques are not commonly used to predict crime, and particularly not used to predict cases of IPV. TabNet, a more novel deep learning approach to classification of tabular data, performed slightly better than multilayer perceptron, but still failed to outperform the ensemble and boosted methods. Overall, all models were found to return high F1 scores, suggesting good precision and recall, but moderate AUC, suggesting wavering sensitivity or specificity, for the FRA data. In the case of classifying IPV, it would be preferable to overestimate the number of women suffering IPV than to underestimate and potentially miss a critical victim who could have been offered help. Hence, high specificity would be more desirable than high sensitivity. While AUC indicates the balance between the two, and was chosen to provide an overall insight into model performance, future work should focus on finding the model that displays the highest specificity so that all women suffering IPV have the potential to be identified.

Upon commencement of this research, it was wondered if a different model would be identified as the most appropriate for the two datasets due to the difference in their origins. This may be true, as the stacked classifier performed optimally when modelled on the NCVS data despite not appearing in the top metrics produced by the models on the FRA data (Table 2). Like on the FRA data, no one model returned the highest metrics for the NCVS data. It is noted that overall accuracy and F1 are much lower when modelled on the NCVS data than the FRA data. In fact, Random Forest on the FRA data returned an accuracy 8% higher than the top performing model on the NCVS data, and average accuracy for all models returned by the two datasets was 12% higher when

[9]https://time.com/5942127/russia-domestic-violence-women/

modelled on the FRA data compared to the NCVS data. It is thought that this may be attributed to the difference in the amount of data between the two datasets. While both contained similar training points following cleansing and transformation, the FRA training set contained a much higher number of features compared to the NCVS training set, 134 versus 16 respectively. This difference may explain why the classification models returned preferable metrics when detecting IPV from the FRA test sets. To determine if this difference is due to the differing amounts of data and if the most appropriate model differs between the two locations, further work would be required where the same variables and volumes of data are collected in both Europe and the US.

From the literature review, it was found that the same few models were applied to data describing violence against women. It was wondered if by applying many different models would a more appropriate one that had not previously been applied for the identification of abuse be found. For this reason, nine models were chosen, including novel applications, models that appear in the pertinent literature and a mixture of algorithm types. Despite these novel applications, this study found Random Forest to return optimal results when predicting violence against women, similar to the results reported by Amusa *et al.* (2020) and Hossain *et al.* (2021). Amusa *et al.* (2020) found that Random Forest returned an optimal AUC of 76% but an accuracy of 53% when predicting married women's risk of IPV. This study found that Random Forest returned a higher accuracy of 88% and 80% for the FRA and NCVS data respectively, but only AUC of 55% and 68%, considerably less than that achieved by Amusa *et al.* (2020). Both datasets of this study modelled on Decision Tree outperformed the accuracy reported for the same model by Amusa *et al.* (2020). Hossain *et al.* (2021) reports accuracies of 77% and 62% for Random Forest and Naïve Bayes when predicting family violence including IPV. When modelled on the FRA data, this study returned higher accuracies than those reported by Hossain *et al.* (2021) for Random Forest and Naïve Bayes, and for Random Forest modelled on the NCVS data. However, this study found that Naïve Bayes returned a lower accuracy than Hossain *et al.* (2021) when modelled on the NCVS data. Reza *et al.* (2021) reported that XGBoost performed best from a selection of models tested, returning a R^2 value of 99%. While this research did not report R^2 as a metric, deeming it inappropriate on non-regression based machine learning methods, XGBoost was also found to return the second best accuracy and the joint highest F1 score when modelled on the FRA data, matching the F1 score returned by Random Forest for the same data. Overall, the results of this study validate the results returned by previous studies, confirming that the models applied in the pertinent literature including Random Forest and XGBoost are appropriate for the task. However, this study also found the novel application of a stacking classifier comprised of Random Forest, Support Vector Machine, Naïve Bayes and XGBoost, as well as Light Gradient Boosting Machine and Support Vector Machine can return satisfactory results for different metrics. It appears that, like in online detection studies, Support Vector Machine can perform well at detecting abuse within the community (Rodríguez-Sánchez *et al.*, 2020; García-Díaz *et al.*, 2021). It is also noted that, particularly when modelled on the FRA data, the methods applied returned accuracies that surpass those reported in the literature. This is likely due to the amount of data passed to the models, as this research applied a greater number of features than other studies. The deep learning methods did not perform as hoped, returning some of the lowest metrics. While deep learning can be a powerful classifier, these models tend to be most effective on unstructured data, such as image data or text data for natural language processing (Ryan, 2021). The data analysed in this study were highly structured, hence these results should have been anticipated.

Based on the results returned by the models applied in this study, it would be recommended to identify potential victims of IPV within the community or for future studies, that a model such a Random Forest, XGBoost, Support Vector Machine or a stacking classifier encompassing these methods be chosen, dependent on the metric of interest. Light Gradient Boosting Machine would also be recommended as it returned good performance in this study, and along with XGBoost, showcases the potential of boosted methods over others. This is also supported by the results returned by other studies applying similar methods (Amusa *et al.*, 2020; Hossain *et al.*, 2021; Reza *et al.*, 2021). It would be recommended that a large number of features be sourced and passed to the models to improve performance. Deep learning techniques would not be recommended for community use at this moment but future work should focus on identifying a deep learning technique that works well with structured data, or by optimising a multilayer perceptron to return better results for these data. While an effort was made to test both novel methods and methods appearing in the literature, some methods were not tested, such as regression. Regression was not modelled in this study as the metrics returned by regression are not directly comparable to accuracy, AUC and F1 score. Regression has been found to perform well in the published literature and this research could be critiqued for omitting this method. Future work could address this by applying regression based methods to the data modelled in this study and comparing the results to those of Raj *et al.* (2021) and Dehingia *et al.* (2022).

7.3 Feature Importance

Permutation feature importance is reported in order to determine the features with the most predictive power in the selected model. Interestingly, the features found to be the most important in the FRA data tended to be related to the victim and her relationship with her partner, while the two most important features identified by Random Forest on the NCVS data related to the number of people living with the victim, and were not directly related to herself or her partner (Figure 4, Figure 5).

Impurity based feature importance was also completed for the models. The decision was made to report the permutation feature importance graphs, as impurity based methods can inflate the importance of variables depending on how they were transformed. Impurity based feature importance tends to mistake features with higher cardinality or that have been overfit for important ones (Masís, 2021). However, there tended to be overlap between the top features identified by both methods of feature importance. Due to the appearance of the same features between permutation and impurity based feature importance, it would be recommended that if data were to be collected within the community, there should be a focus on collecting similar features to those identified as important by this study. Interestingly in Figure 4, the country Finland was identified as the 15^{th} feature with the most predictive power. Countries such as Greece, Netherlands and Slovenia were also returned as top features by impurity based feature importance when modelled on the FRA data. Hence, it would be recommended that geographical regions also be collected should these models be considered for community use. Further analysis would be required to determine if smaller geographical regions would contribute as much predictive power as countries, however it is thought that geography plays an important role in whether a woman suffers abuse or not due to the results of the exploratory analysis of the FRA data. McFarlane *et al.* (2016) found that the amount of support available to a woman affected the likelihood of her returning to a shelter for domestic

abuse victims. It would be logical to assume that like country, smaller geographical areas would be found to provide predictive power in modelling dependent on the local policies and support available within that area.

Some features such as race and occupation were not often identified by the models as being important features. It is thought that this may be due to how the data were aggregated, both upon collection and at the data cleansing and processing stage. Data aggregation can be necessary, particularly when resources are limited or when the collection surveys are long. However, information can be quickly lost by only recording broad groupings of data, such as race and occupation. This study should be critiqued for re-grouping multiple races into broader groupings for the NCVS data. Race can vary widely and limiting this to two or three broad groupings can be very unhelpful, hiding the fact that this can be a very diverse group with a potentially varied rate of IPV suffered. Future collections of such data should allow for greater nuance in race, ethnicity and occupation recorded. As well, data on cultural receptiveness should be collected to determine if there is a relationship between culture and IPV. Differences were found in the role of economic independence in the prediction of IPV between South America and Africa (Borraz and Munyo, 2020; Alesina *et al.*, 2021). Further data collection and analysis without data aggregation would be required to determine if this difference is due to cultural nuances or other factors. As different datasets were used, it is difficult to compare the results of this study to the features identified as important in other studies, such as Dehingia *et al.* (2022) and Borraz and Munyo (2020). However, through more robust feature analysis techniques such as permutation feature importance in comparison to regression, novel features were identified as playing a role in the rate of IPV suffered by women.

Feature selection was not carried out as there were not a high number of features prior to transformation. Relationships between the variables were determined within the data exploration phase. No high correlations or important interactions were found. For these reasons, feature selection was not deemed to be appropriate. If more data and unique variables were to be collected, this should be considered.

7.4 Error Analysis

Unexpected and interesting results were returned by the error analysis. Table 3 shows that when the number of features were reduced and input into Random Forest, the variance in accuracy or the relative error of the model remained stable. However, the accuracy returned was reduced by almost 10%. This supports the earlier theory that the reduced accuracies returned by the models when the smaller NCVS training set was the input may be due to the reduced amount of features in comparison to the FRA training set. In contrast to reducing the number of features, when the number of training points were reduced, the average accuracy remained relatively stable while the error returned was almost doubled. When the number of training points were increased, it was found that the mean accuracy returned increased by a significant amount whilst the error returned was reduced to an almost negligible value. However, as the training points were increased by sampling the data with replacement, it's possible that accuracy and error were improved due to the models being trained on an increased amount of the same data. Future work should determine if this is the case by obtaining an increased amount of data describing IPV against women and increasing the number of training points without replacement.

The results obtained by error analysis provide important insights for future classification studies detecting IPV. While the results show that by collecting more data, accuracy

may be improved, there is good evidence to suggest that a reduction of features does not destabilise error. The results also suggest that a reduction of training points does not destabilise average accuracy. Further research into error analysis would be required to determine if this is the case and the effect on additional models should be tested. However, from this research, it would be recommended that if time or resources are a constraint as they often are in research, a reduced number of features or training points could be accepted whilst maintaining adequate results dependent on the objectives. From a thorough literature review, it was noted that the error returned by models was not reported. This is a novel aspect of research in the domain of classifying IPV against women, and further work should be carried out to confirm the results. However, preliminary analysis suggests that model performance could be maintained despite reduced input, which could save future research time and budgetary resources.

7.5 Potential Benefits and Applications

Aside from classification research in the domain of violence against women, a number of areas where this research may be useful to stakeholders has been identified. Upon undertaking this research, the published reports from a number of women's charities such as Women's Aid[10] and Aoibhneas[11] were consulted. As far back as 2008, Aoibhneas suggested that the rate of crime should not be predicted to determine the number of arrests that will be made, but should be predicted to ensure the safety of women (Murphy and McDonnell, 2008). This research aimed to keep the safety of women in mind with every decision made. Aoibhneas states that the purpose of determining the risk of domestic violence is to prevent further violence and ultimately death. If the recommendations made by this study were to be implemented by a charity such as Women's Aid or Aoibhneas, it is thought that risks could be quickly and accurately identified to prevent further violence and femicide. In fact, it is believed that this tool could be used by more than just women's charities. Services that deal with both victims of abuse and the perpetrators should be equipped to assess this risk, including healthcare providers, women's refuges, domestic violence charities, An Garda Síochána and court, probation and social services. This could be done by utilising any of the high performing models identified by this study, to detect the risk of IPV following extensive collection of data within the area. According to Murphy and McDonnell (2008), an important way to increase the safety of women and children experiencing domestic violence is to conduct risk assessments and use the assessments to help victims understand that their perception of risk is vital to keeping them safe. It has been found that a woman's perception of risk is a big predictor of re-assault (Murphy and McDonnell, 2008). If the recommendations made by this were study were to be implemented, not only could women be quickly and discretely identified as at risk of IPV, but they could be educated on the relative risks that led to their identification to prevent re-assault. However, if these models were to be adopted for community use, it is noted that not only would it be required to explain to stakeholders how the models work to return a result and how to use them, but that a working relationship would need to be established. It would require a lot of trust and ethical consideration for a women's charity or medical body to partake in a program that utilises any of the models tested in this study, as they can be difficult to interpret without training in such techniques. There may be apprehension surrounding the security and safety of the women identified by such

[10]https://www.womensaid.ie/about/policy/publications.html
[11]https://aoibhneas.ie/resources/annual-report-2020/

models, and care should be taken to ensure these models return the most accurate results whilst maintaining victims' anonymity. Given the restraints of the use of private data, building such a relationship was out of the scope of this study. Future work should collaborate with stakeholders so that they can be educated on the benefits of classification and provide their input so that these models may be used to help victims.

There is a clear need for continued research into ways to solve IPV and wider abuse of women in Ireland. In July 2022, the Allied Irish Bank (AIB) announced a plan to remove cash services from 70 of its nationwide banks, causing Women's Aid and the National Council of Ireland to respond that this would negatively affect women under the control of their partners suffering financial abuse[12]. Women who have greater financial independence may suffer less domestic violence according to Borraz and Munyo (2020), and by maintaining cash facilities in banks, women suffering abuse may have greater access to money that cannot be as easily tracked and controlled. AIB ultimately reversed their decision. Studies such as this one, which have the ability to estimate the rate of IPV in a particular area, could facilitate better decision making by businesses such as AIB and help to avoid public backlash. If the resources identified by this study were made available to the community, banks could assess if it is ethical to remove cash facilities in a particular branch first. In May 2022, the Department of Justice announced that an information guide for recognising patients suffering domestic abuse would be made available to General Practitioners (GPs)[13]. It was noted that a coordination of services would be required in order to identify and protect such victims. The guide itself contains general statistics and risk factors, and advice on how GPs should address this sensitive topic with their patients[14]. However, the risk factors cited are broad and generalised, while this study identified that risks may be area specific, seen in the difference of risks identified between the European and US data. While this guide is useful and a step in the right direction for victims, a more personalised approach would be required to identify the victims that may otherwise be overlooked. The research carried out by this project could be used to inform future guides. The HSE is under statutory duty to identify children at risk due to the Child Care Act 1991, but not women vulnerable to violence, despite the WHO recognising violence against women as a major public health issue[15]. Of course there is a fundamental requirement for a thorough ethical review around the potential use of such a tool. Victims of IPV require the utmost security, and no victim should be identified if it places them at further risk. On balance, there is a need in society and healthcare to utilise all available resources to identify the most vulnerable in society and it is believed that this research has identified multiple tools that could be put to use.

8 Conclusion and Future Work

In this study, multiple knowledge gaps were addressed by testing many classification techniques to determine the most appropriate model for identifying women at risk of IPV. The research question was twofold: to what extent could machine and deep learning techniques identify female victims of IPV using two different, geographically distinct datasets,

[12]https://www.thejournal.ie/cashless-branches-aib-womens-aid-financial-abuse-5828953-Jul2022/

[13]https://www.icgpnews.ie/press-release-new-guide-on-management-of-domestic-violence-abuse-launched/

[14]https://www.icgpnews.ie/wp-content/uploads/2022/05/ICGP-Domestic-Violence-QRG-Summary.pdf

[15]See Footnote 1

and could an appropriate predictive model and the factors influencing it be identified? These questions were addressed by applying nine models to two datasets describing IPV against women. This was to determine if the models appearing in the current literature were the most appropriate for the task, or if a novel application performed better. It was found that multiple techniques can accurately identify women at risk with varying amounts of data. Risk assessments are an important way of protecting women from violence (Murphy and McDonnell, 2008). There is potential for the techniques applied in this study to be implemented as a more personalised risk assessment tool or to improve current methods of risk assessment.

Both previously applied and novel models including Random Forest, XGBoost, Light Gradient Boosting Machine, Support Vector Machine and a stacking classifier were found to be particularly appropriate for classifying victims whilst returning stable levels of error. From this, a number of recommendations for both future research and potential implementation of a tool for community use were made. These recommendations can be summarised as applying techniques that consistently return high accuracy, AUC or F1 score while maintaining stable variance despite reduced amounts of data. It would also be recommended that data describing not only the woman's relationship with her partner but her household situation, be modelled. It is believed that the error analysis undertaken in this research is of significant importance to future classification studies in this domain. Error analysis was a large identified gap in the pertinent literature, despite the potential value that the results of this analysis returned. If resources such as time and funds are limited, then the amount of data collected could be limited without sacrificing accuracy or increasing error, as was found in this study. Deep learning techniques would not be recommended based on this research, but these techniques are constantly evolving and improving and novel deep learning techniques should be considered in the future.

There are many areas within the domain of classifying IPV that still need to be addressed. Future work should focus on applying the same models to smaller, more localised geographical areas. There should be a focus on refining the results for each metric returned by each model. There is potential to further reduce the error and given more time, greater research into the error returned by models would have been conducted. However, it is noted that no study in the domain of classifying violence against women has assessed model error, hence this study provides novel insights into the capabilities of these models on such data. The data used in this study could be considered a limiting factor. It was found upon undertaking this research that few bodies publish data describing women suffering abuse. This is sensible due to the sensitive nature of such data, but does impede advancement within this domain. Once data were acquired, it was found to be highly aggregated, at a geographical and personal level. Data aggregation, while often a requirement due to limited resources, can hide what is happening to a particular population of people. When disaggregation is impossible due to the method by which data were collected, information that could be useful to models and analysis is lost. While feature analysis was carried out to determine the most important predictors in IPV across the different models, this analysis was limited to the data available. Experts and survivors should be engaged to inform what factors should be collected and analysed in future studies. For the purpose of this study, an assumption was made that IPV was limited to physical or sexual abuse, and that only women were victims. This definition was chosen due to the available data, and to avoid a definition that was too broad or too specific. However, Aoibhneas state that the majority of abuse suffered by their clients

is emotional[16]. As well, IPV is not gender specific and people of all genders can be victims too. The same research should be replicated with data describing IPV for both women and men, and include emotional abuse in the target variable to truly determine the effectiveness of such a tool within the community.

In conclusion, previous to this study there was a focus on detecting abuse towards women online or using the same techniques to detect IPV using few variables. Eight identified knowledge gaps were addressed by answering the research question through the six objectives. This research shows that both novel applications and models appearing in the current literature may be the most appropriate for identifying IPV. Deep learning did not perform as was hoped. There is still a capacity to improve research through better collection of data, which may improve the error returned by models. This study applied novel techniques to data that had yet to be analysed in such a way, and highlights the potential technology has to benefit victims of abuse. Within classification, there is a need to identify the best model rather than blindly picking one for research. There is potential for this study to aid future research by allowing for the most appropriate model and features for predicting IPV to be selected whilst limiting error returned. It is hoped that this work will contribute to identifying women suffering IPV within the community, so that they can be offered the help they require and deserve.

9 Acknowledgements

This research would not be possible without the help of the staff at NCI who taught me over the last three years. A special thanks to the European Union Agency for Fundamental Rights for allowing me access to their data, and to the UK Data Service for helping me throughout the application process. I would particularly like to thank my supervisor, Dr. Vladimir Milosavljevic for all of his help and advice, and for encouraging me to push my research further.

Finally, I owe special thanks to my parents, Andy and Louise, friends, particularly Aoife, Gráinne and Tom, colleagues, my beloved Luna, and especially my partner, John. Without them, I would not have made it past the first semester of the Higher Diploma. I am forever grateful and indebted.

References

Alesina, A., Brioschi, B. and La Ferrara, E. (2021) 'Violence against women: A cross-cultural analysis for africa', *Economica*, 88(349), pp. 70–104.

Amusa, L. B., Bengesai, A. V. and Khan, H. T. (2020) 'Predicting the vulnerability of women to intimate partner violence in south africa: Evidence from tree-based machine learning techniques', *Journal of interpersonal violence*, pp. 1–18.

Anderson, E. J., Krause, K. C., Meyer Krause, C., Welter, A., McClelland, D. J., Garcia, D. O., Ernst, K., Lopez, E. C. and Koss, M. P. (2021) 'Web-based and mhealth interventions for intimate partner violence victimization prevention: a systematic review', *Trauma, Violence, & Abuse*, 22(4), pp. 870–884.

[16]https://aoibhneas.ie/wp-content/uploads/2021/10/Aoibhneas-Annual-Report-2020-Final-1.pdf

Arik, S. Ö. and Pfister, T. (2021) Tabnet: Attentive interpretable tabular learning, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 35, pp. 6679–6687.

Baek, M. S., Park, W., Park, J., Jang, K.-H. and Lee, Y.-T. (2021) 'Smart policing technique with crime type and risk score prediction based on machine learning for early awareness of risk situation', *IEEE Access*, 9, pp. 131906–131915.

Batista, G. E., Bazzan, A. L., Monard, M. C. *et al.* (2003) Balancing training data for automated annotation of keywords: a case study., *in* 'WOB', pp. 10–18.

Borraz, F. and Munyo, I. (2020) 'Conditional cash transfers, women's income and domestic violence', *International Review of Applied Economics*, 34(1), pp. 115–125.

Castorena, C., Abundez, I., Alejo, R., Granda-Gutiérrez, E., Rendón, E. and Villegas, O. (2021) 'Deep neural network for gender-based violence detection on twitter messages', *Mathematics*, 9(8), p. 807.

Chen, I. Y., Alsentzer, E., Park, H., Thomas, R., Gosangi, B., Gujrathi, R. and Khurana, B. (2020) Intimate partner violence and injury prediction from radiology reports, *in* 'BIOCOMPUTING 2021: Proceedings of the Pacific Symposium', World Scientific, pp. 55–66.

Chen, T. and Guestrin, C. (2016) Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', pp. 785–794.

Dehingia, N., Dey, A., McDougal, L., McAuley, J., Singh, A. and Raj, A. (2022) 'Help seeking behavior by women experiencing intimate partner violence in india: A machine learning approach to identifying risk factors', *PloS one*, 17(2), p. e0262538.

García-Díaz, J., Cánovas-García, M., Colomo-Palacios, R. and Valencia-García, R. (2021) 'Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings', *Future Generation Computer Systems*, 114(52), pp. 506–518.

Gracia, E. and Merlo, J. (2016) 'Intimate partner violence against women and the nordic paradox', *Social Science & Medicine*, 157, pp. 27–30.

He, J. and Zheng, H. (2021) 'Prediction of crime rate in urban neighborhoods based on machine learning', *Engineering Applications of Artificial Intelligence*, 106, p. 104460.

Hossain, M., Abdulla, F., Rahman, A., Khan, H. T. *et al.* (2022) 'Prevalence and determinants of wife-beating in bangladesh: evidence from a nationwide survey', *BMC psychiatry*, 22(1), pp. 1–13.

Hossain, M., Asadullah, M., Rahaman, A., Miah, M., Hasan, M., Paul, T., Hossain, M. *et al.* (2021) 'Prediction on domestic violence in bangladesh during the covid-19 outbreak using machine learning methods', *Applied System Innovation*, 4(4), p. 77.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017) 'Lightgbm: A highly efficient gradient boosting decision tree', *Advances in neural information processing systems*, 30.

Masís, S. (2021) *Interpretable Machine Learning with Python: Learn to build interpretable high-performance models with hands-on real-world examples*, Packt Publishing Ltd.

McFarlane, J., Pennings, J., Liu, F., Gilroy, H., Nava, A., Maddoux, J., Montalvo-Liendo, N. and Paulson, R. (2016) 'Predicting abused women with children who return to a shelter: Development and use of a rapid assessment triage tool', *Violence against women*, 22(2), pp. 189–205.

Murphy, C. and McDonnell, N. (2008) 'Escalating violence: How to assess and respond to risk. a review of international experience'.

Pace, L. (2012) *Beginning R: An introduction to statistical programming*, Apress.

Raj, A., Dehingia, N., Singh, A., McAuley, J. and McDougal, L. (2021) 'Machine learning analysis of non-marital sexual violence in india', *EClinicalMedicine*, 39, p. 101046.

Rawlings, S. and Siddique, Z. (2020) 'Domestic violence and child mortality in the developing world', *Oxford bulletin of economics and statistics*, 82(4), pp. 723–750.

Reza, R., Mannan, F., Barua, D., Islam, S., Khan, N. and Mahmud, S. (2021) Developing a machine learning based support system for mitigating the suppression against women and children, *in* '2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)', IEEE, pp. 1–6.

Rodríguez, D. A., Díaz-Ramírez, A., Miranda-Vega, J. E., Trujillo, L. and Mejía-Alvarez, P. (2021) 'A systematic review of computer science solutions for addressing violence against women and children', *IEEE Access*.

Rodríguez-Rodríguez, I., Rodríguez, J., Pardo-Quiles, D., Heras-González, P. and Chatzigiannakis, I. (2020) 'Modeling and forecasting gender-based violence through machine learning techniques', *Applied Sciences*, 10(22), p. 8244.

Rodríguez-Sánchez, F., Carrillo-de Albornoz, J. and Plaza, L. (2020) 'Automatic classification of sexism in social networks: An empirical study on twitter data.', *IEEE Access*, 8, pp. 219563–219576.

Ryan, M. (2021) *Deep learning with structured data*, Manning Publications Co.

Sáez, G., Ruiz, M. J., Delclós-López, G., Expósito, F. and Fernández-Artamendi, S. (2020) 'The effect of prescription drugs and alcohol consumption on intimate partner violence victim blaming', *International journal of environmental research and public health*, 17(13), p. 4747.

Safat, W., Asghar, S. and Gillani, S. A. (2021) 'Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques', *IEEE Access*, 9, pp. 70080–70094.

Subramani, S., Wang, H., Vu, H. and Li, G. (2018) 'Domestic violence crisis identification from facebook posts based on deep learning', *IEEE Access*, 6, pp. 54075–54085.

van Gelder, N., Peterman, A., Potts, A., O'Donnell, M., Thompson, K., Shah, N. and Oertelt-Prigione, S. (2020) 'Covid-19: Reducing the risk of infection might increase the risk of intimate partner violence', *EClinicalMedicine*, 21.

Xue, J., Chen, J., Chen, C., Hu, R. and Zhu, T. (2020) 'The hidden pandemic of family violence during covid-19: unsupervised learning of tweets.', *Journal of medical Internet research*, 22(11), p. 24361.

Zara, G. and Gino, S. (2018) 'Intimate partner violence and its escalation into femicide. frailty thy name is "violence against women"', *Frontiers in psychology*, 9, p. 1777.

Zhang, X., Liu, L., Lan, M., Song, G., Xiao, L. and Chen, J. (2022) 'Interpretable machine learning models for crime prediction', *Computers, Environment and Urban Systems*, 94, p. 101789.

Zhang, X., Liu, L., Xiao, L. and Ji, J. (2020) 'Comparison of machine learning algorithms for predicting crime hotspots', *IEEE Access*, 8, pp. 181302–181310.