**Accepted version 20 August 2020**

**Predicting the vulnerability of women to intimate partner violence in South Africa: Evidence from tree-based machine learning techniques**

**Lateef B. Amusa1, Annah V. Bengesai2 andHafiz T.A. Khan3**

1Department of Statistics, University of Ilorin, Nigeria

2 College of Law and Management Studies, University of KwaZulu-Natal, 4000, South Africa

3 College of Nursing, Midwifery and Healthcare, University of West London, UK

**Abstract**

Intimate partner violence (IPV) is a pervasive social challenge with severe health and demographic consequences. Global statistics indicate that more than a third of women have experienced IPV at some point in their lives. In South Africa, IPV is considered a significant contributor to the country’s broader problem with violence, and a leading cause of femicide. Consequently, IPV has been the major focus of legislation and research across different disciplines. The present paper aims to contribute to the growing scholarly literature by predicting factors that are associated with the risk of experiencing IPV. We used the 2016 South African Demographic and Health Survey dataset and restricted our analysis to 1816 ever-married women who had complete information on the variables that were used to generate IPV. Prior research has mainly used regression analysis to identify correlates of IPV; however, while regression analysis can test a priori specified effects, it cannot capture unspecified inter-relationship across factors. To address this limitation, we opted for machine learning methods, which identify hidden and complex patterns and relationships in the data. Our results indicate that the fear of the husband is the most critical factor in determining the experience of IPV. In other words, the risk of IPV in South Africa is associated more with the husband or partner’s characteristics than the woman’s. Such models can then be used to develop interventions by different stakeholders such as social workers, policymakers and or other interested partners.

**Keywords:** intimate partner violence, machine learning, decision tree, South Africa

**Introduction**

Intimate partner violence is a pervasive social challenge with severe health and demographic consequences. While IPV can be perpetrated against men or women, research suggests that women are disproportionately affected (World Health Organisation (WHO), 2015; Abramsky et al. 2011; García-Moreno et al. 2005). Global statistics indicate that more than a third of women have experienced IPV at some point in their lives, although there are regional and country-specific differences (WHO, 2015). In South Africa, IPV is considered a significant contributor to the country’s broader problem with violence, which has sparked debates leading to law reforms, and widespread activism (Vetten, 2014; Abrahams et al. 2010). The 2016 South African Demographic and Health Survey indicates that approximately 20% of ever partnered women have experienced physical violence, while 17% have reported emotional violence and 6% sexual violence. Overall, 26% of South African women have experienced some form of IPV at some point in their lives (National Department of Health et al. 2019). IPV is also considered the leading cause of mortality among South African women, and the second-highest burden of disease after HIV and AIDS (Gordon, 2016; Abrahams et al. 2013; Abrahams et al. 2013). As many as 5.6/100 000 women are killed by their intimate partners, a rate which is perhaps the highest in the world (Abrahams et al. 2012).

These statistics are troubling, considering many studies have demonstrated the adverse effects of IPV. Violence against women has been linked to several physical, mental, sexual and reproductive health consequences (WHO, 2015; Abramsky, 2011). Victims of IPV are most likely to report physical injuries, psychological morbidities such as depression and post-traumatic stress disorder, gynaecological disorders, as well as HIV and AIDS (Gibbs et al. 2017; Yaya et al. 2019; Sugg, 2015; Shamu et al. 2011).

Several socio-behavioural factors, most of which are linked to power relations between men and women (Chikhungu et al. 2019; Eriksson & Mazerolle, 2015; Jewkes, 2002) have been associated with IPV. There is substantial evidence that social norms and attitudes that support violence in general (Jewkes et al. 2012; Abrahams et al. 2006) and specifically against women have contributed to the high levels of IPV (Yilmaz, 2018; Eriksson & Mazerolle, 2015). In many patriarchal communities, women are considered as subordinate to men, and violence is seen as a normal way of resolving conflict where societal expectations of culturally assigned gender roles are transgressed (Dekel & Andipatin, 2016; Allen & Devitt, 2012). Thus, permissive attitudes towards gender-based violence not only condone the perpetration of IPV but also influence victims’ responses to violence (Allen & Devitt, 2012). The risk of IPV has also been linked to socio-economic factors. Several studies have demonstrated that women from low socio-economic backgrounds are at higher risks of IPV (Chikhungu et al. 2019; Reichel, 2017; Abramsky 2011), while economic autonomy has been shown to be endogenously associated with IPV (Fakir et al. 2016). In some instances, a correlation between witnessing interparental violence and vulnerability to IPV has been reported (Eriksson & Mazerolle, 2015; Islam et al. 2014). Specific to South Africa, a history of violence, having multiple partners, as well as the level of education, are some of the factors that have been linked to the risk of IPV (Zembe et al. 2015; Jewkes, 2002).

IPV has many contributing factors, and globally, there are several studies which have investigated the effect of these on determining one’s vulnerability to IPV (Chikhungu et al. 2019; Eriksson & Mazerolle, 2015; Abramsky et al. 2011; Jewkes, 2002). However, most of these earlier studies have relied mainly on traditional regression analyses to identify the correlates. While regression analysis can test a priori specified effects, it cannot capture unspecified inter-relationships across factors. Machine learning (ML) methods can address these limitations since they utilise a variety of “statistical, probabilistic and optimisation techniques” to identify hidden and complex patterns and relationships in the data. ML is increasingly being used to create algorithms that have shown to have relatively more predictive reliability than the conventional methods (Uddin et al. 2019; Bisaso et al. 2017). Recently, ML models have been applied in many areas of medicine and public health (Galatzer-Levy, 2017; Lecun et al. 2015; Yang et al. 2009); however, the approach is yet to be adopted by IPV researchers using population-based data. For instance, in our literature search, we came across only two studies that utilised ML models for risk prediction and achieved acceptable predictive accuracies. Ghosh (2007) used classification trees and random forests to predict the vulnerability of ever-married women aged 15-40 to domestic violence incidents in India. In another study where the focus was on IPV perpetration, Petering et al. (2018) used several supervised ML algorithms, including logistic regression, support vector machines, random forests, and neural networks, to build an IPV perpetration triage tool which could be used to identify young people who are at high risk of perpetrating IPV.

The aim of this paper is two-fold: first, to build predictive models that can efficiently classify women based on their likelihood of experiencing IPV; second, to identify significant factors associated with the risk of IPV among these women. We use population-based data to build an IPV vulnerability model. We build tree-based machine learning models, including decision trees (DT), random forests (RF), gradient boosting (GB), as well as the conventional logistic regression (LR) model, to predict the occurrence of IPV. Models developed in this study can be used to effectively identify subgroups of vulnerable women with a high risk of IPV and improve the efficiency of interventions for such women.

**Data and Methods**

We used data from the 2016 South African Demographic and Health Survey (SADHS), which was undertaken by Statistics South Africa, in collaboration with the South African Medical Research Council (SAMRC). The SADHS is a cross-sectional and nationally representative survey of the health and demographic indicators for women aged 15-49. Data were collected from all nine provinces using a Master Sampling Frame (MSF) which was delineated using the Enumeration Areas (EAs) of the 2011 South African Census. These provinces were further stratified into the urban, farm and traditional areas, resulting in 26 sampling strata. From these, 750 primary sampling units were designated; 468 in urban areas, 224 and 59 in traditional and farm areas, respectively (National Department of Health et al. 2019).

 **Study Participants**

Overall, 8720 women aged 18 years and older, who were eligible for the domestic violence module were selected for individual interviews. Although 2442 women agreed to be interviewed for domestic violence, we only included ever-married women who also had complete information used to generate the IPV variable. Our final analytic sample was made up of 1816 women.

**Outcome Variable**

We defined the principal outcome variable, IPV, as any form of physical, sexual, and emotional abuse perpetrated against women by their partners. Each of the three forms of abuse (physical, sexual, and emotional) were determined by response to a set of questions (see Table 1) as measured by the SADHS (National Department of Health et al. 2019).

**Table 1: Set of questions associated with IPV in the dataset**

|  |  |
| --- | --- |
| S/N | Items |
| 1 | Ever been pushed, shaken or had something thrown at by husband/partner |
| 2 | Ever been kicked, dragged or beaten up by husband/partner |
| 3 | Ever been choked up or attempted to get burnt by husband/partner |
| 4 | Ever been threatened with knife/gun or other weapons by husband/partner |
| 5 | Ever had arm being twisted or hair pulled by husband/partner |
| 6 | Ever been slapped by husband/partner |
| 7 | Ever been punched with fist or hit with something harmful by husband/partner  |
| 8 | Ever been physically forced into unwanted sex by husband/partner |
| 9 | Ever been forced with threats to perform sexual acts you did not want to |
| 10 | Ever been physically forced to perform sexual acts you did not want to |
| 11 | Ever been humiliated by husband/partner |
| 12 | Ever been threatened with harm by husband/partner |
| 13 | Ever been insulted or made to feel bad by husband/partner |

A “yes” response to any of questions 1-7, 8-10, and 11-13 constituted physical violence, sexual violence, and emotional violence, respectively. For our analysis, we considered any woman who responded yes to at least one of the 13 items as being a victim of IPV.

We selected risk factors of IPV based on causal assumption derived from subject matter knowledge and literature review, which comprised respondents’ demographic, social, economic, union, and household characteristics (Zembe et al. 2015; Johnston & Naved, 2008; Ghosh, 2007; García-Moreno et al. 2005; Jewkes, 2002; Krug et al. 2002). These included current age, age at first cohabitation/marriage, spousal age difference, marital duration, place of residence, ethnicity, level of education for both the respondent as well as the partner, household wealth index, religion, number of household members, sex of household head, employment status, number of living children, partner’s alcohol and drug usage, history of abuse, empowerment variables, and variables of attitude to wife-beating. A full description of these variables can be found in Table 2.

**Statistical Analysis**

We first explored the associations between IPV and baseline characteristics. We utilised chi-square tests categorical variables, while t-tests were used to determine the relationship between IPV and the continuous variables (Table 2). We reported two-sided p-values, and statistical significance was determined at P < 0.05. The analysis was done in Stata v.14 (Stata Corp, 2015). We developed models for IPV prediction using four ML algorithms: decision trees, using the classification and regression trees (CART) algorithm (Breaiman et al. 1984), random forests (Breiman, 2001), gradient boosting (Friedman et al. 2002; Friedman, 2000), and logistic regression (Hosmer et al. 2013). Tree building algorithms offers a variety of flexible methods which built a common framework for CART and ensembles such as random forests and gradient boosting. Random forest (RF) and gradient boosting (GB) algorithms were chosen for their resiliency to overfitting, relative ease in the implementation, and general acceptance in the machine learning community (Taylor et al. 2018). Decision trees (DT) was selected due to its intuitiveness and easily interpretable pictorial evidence, despite its ability to reveal complex non-linear associations. Logistic regression (LR), which is the commonly used discriminatory model in applied studies, was chosen as a baseline comparison.

The implementation of tree-based algorithms is not a black box. Tree-based methods generally involve stratifying the predictor space into a number of simple regions. Predictions on a given observation are typically made using the mean or mode of the observations in the region it belongs to. In its simplest form, an equivalent tree-based method, which grows only one tree, is termed a decision tree. A decision tree is a simple structure that represents how we make decisions, like an if-this-then-that game. The process of growing a decision tree (using the CART algorithm) involves taking a recursive splitting approach to tree building. From the first split (known as the root node), which contains the entire data for model fitting, the splitting process is repeated sequentially, while achieving a top-down tree structure and the most homogeneous subregions at each variable and cutpoint. The splitting is repeated along the child nodes (tree branches) until a terminal node (leaf) is reached.

Though simple and highly intuitive, a single decision tree does not have prediction power as high as some of the other regression and machine learning algorithms. Random forests and gradient boosting models, also known as ensemble models, build upon decision trees by aggregating many trees to construct more powerful prediction models. Random forests involve simultaneously building simple decision trees using subsets of randomly selected data (with replacement). Further, each tree uses only a subset of the predictors. Gradient boosting, on the other hand, does not involve random sampling of data; the trees are grown sequentially: previously grown trees are used to grow each tree.

Seven variables had missing data with a percentage of missingness ranging from 0.2% - 5.7%. Thus, before applying LR, missing values were imputed using tree imputation with surrogate splitting rules (Borgoni, & Berrington, 2013). The tree-based methods, including decision trees, gradient boosting, and random forest can handle missing values internally.

We developed the models using the full set of variables described in Table 2. For LR, the least absolute shrinkage and selection operator (LASSO) variable selection technique was used in conjunction with the Akaike Information Criteria (AIC) to include only variables that avoid over-fitting and maximises the potential usefulness of the final model. For performance comparison, we primarily reported for each algorithm, the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. Additional reported statistics included balanced accuracy, sensitivity (recall), specificity, precision, and F1 score (Mitra, 2009). To avoid overfitting, we incorporated a three-way data split: all models were trained, validated, and tested on a ratio of 60:20:20 random partition of the data. The test set was chosen for evaluating performance on unseen data. Where applicable, hyperparameters were tuned and optimised through stratified cross-validation and exhaustive grid searches within the training data.

Finally, in line with the study objectives, it is essential to construct an interpretable classification model (Ferreira et al. 2012). Although we are ultimately interested in the model with best predictive performance, discriminative features or characteristics for IPV can be interpretable with by algorithms such as the DT or LR. Here, we focus on the DT model. Thus, interesting splitting rules were generated using IF-THEN statements and displayed in a decision tree for decision making. To avoid overfitting of the training data that results from very large trees, we introduced the stopping rule of a fewer than 10% of the training sample in order to limit the tree size (Lemon et al. 2003). We performed all predictive modelling in the environment of SAS enterprise miner 14.2 software.

**Results**

Table 2 shows the descriptive statistics and univariate analysis of selected characteristics stratified by the outcome variable, IPV. Women who experienced IPV accounted for 21.9% of the total sample. Our initial analysis identified risk factors that give useful insights into IPV vulnerability. Risk factors that were significantly associated (P < 0.05) with IPV include ethnicity, history of spousal abuse of respondent’s mother, household wealth index, attitude to wife-beating, whether husband/partner takes drugs, whether husband/partner drinks alcohol, whether the respondent is afraid of husband/partner, husband/partner’s education, whether respondent decides her health care, and whether respondent decides large household purchase.

**Table 2:** Descriptive characteristics of study respondents

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label  |  | IPV: YesN=398 (21.9%) | IPV: NoN=1418 (78.1%) | P-Value |
|  | **Sociodemographic/economic characteristics** |  |  |  |
| **V102**  | Place of residence |  |  | 0.073 |
|  | *Urban* | 276 (69.3) | 1056 (74.5) |  |
|  | *Rural* | 122 (30.7) | 361 (25.5) |  |
| **V106**  | Educational level |  |  | 0.058 |
|  | *No education* | 13 (3.3) | 37 (2.6) |  |
|  | *Primary* | 46 (11.6) | 128 (9.0) |  |
|  | *Secondary* | 295 (74.1) | 999(70.5) |  |
|  | *Higher* |  44(11.1) | 253(17.9) |  |
| **V131** | Ethnicity |  |  | **0.010** |
|  | *Black/African* | 330 (82.9) | 1152 (81.2) |  |
|  | *White* | 8 (2.0) | 98 (6.9) |  |
|  | *Coloured* | 51 (12.8) | 144 (10.2) |  |
|  | *India/Asian* | 5 (1.3) | 24 (1.7) |  |
|  | *Other* | 4 (1.0) | 0 (0.0) |  |
| **V151** | Sex of household head |  |  | 0.686 |
|  | *Male* | 304 (76.4) | 1102 (77.7) |  |
|  |  *Female* | 94 (23.6) | 316 (22.3) |  |
| **V714** | Currently working |  |  | 0.103 |
|  |  *No* | 244 (61.3) | 786 (55.4) |  |
|  | *Yes* | 154 (38.7) | 632 (44.6) |  |
| **Age\_diff** | Age difference |  |  | 0.751 |
|  | *Same age or wife older* | 64 (16.0) | 215 (15.2) |  |
|  | *Husband older* | 335 (84.0) | 1203 (84.8) |  |
| **DD121** | History of spousal abuse of respondent’s mother |  |  | **<0.001** |
|  |  *No* | 272 (72.3) | 1204 (88.3) |  |
|  | *Yes* | 104 (27.7) | 160 (11.7) |  |
| **V190** | Wealth Index |  |  | **<0.001** |
|  | *Poor* | 194 (48.7) | 486 (34.3) |  |
|  | *Average* | 76 (19.1) | 310 (21.9) |  |
|  | *Rich* | 128 (32.2) | 622 (43.9) |  |
| **Edu\_diff** | Spousal Educational difference |  |  | 0.962 |
|  |  *Wife equally or less educated* |  340 (88.8) | 1220 (88.5) |  |
|  |  *Husband more educated* |  43(11.2) | 158 (11.5) |  |
| **Duration** | Marital duration |  |  | 0.583 |
|  | *Short* |  116 (29.1) | 407 (28.7) |  |
|  |  *Medium* |  169 (42.4) | 559 (39.4) |  |
|  | *Long* |  114 (28.6) | 452 (31.9)  |  |
| **Attitude** | Attitude to wife-beating |  |  | **<0.001** |
|  |  *Justified* | 335 (84.2) | 1378 (97.2) |  |
|  |  *Unjustified* | 63 (15.8) | 40 (2.8) |  |
| **V012** | Respondent’s current age | 34.1 ± 7.8 | 35.1 ± 7.7 | 0.097 |
| **V218** | Number of living children | 2.2 ± 1.5 | 2.2 ± 1.3 | 0.751 |
| **V511** | Age at first cohabitation | 23.4 ± 6.1 | 24.2 ± 6.3 | 0.088 |
| **V136** | Number of household members | 4.5 ± 2.5 | 4.1 ± 2.1 | 0.114 |
|  | **Husband/Partner’s characteristics** |  |  |  |
| **SS1512a** | Husband/Partner takes drugs |  |  | **<0.001** |
|  |  *No* | 372 (94.4) | 1394 (98.4) |  |
|  | *Yes* | 22 (5.6) | 22 (1.6) |  |
| **DD129** | Respondent is afraid of husband/partner |  |  | **<0.001** |
|  |  *No* | 241 (60.6) | 1245 (87.9) |  |
|  | *Yes* | 157 (39.4) | 172 (12.1) |  |
| **DD113** | Husband/Partner drinks alcohol |  |  | **<0.001** |
|  |  *No* | 169 (42.5) | 949(67.1) |  |
|  | *Yes* | 229 (57.5) | 466 (32.9) |  |
| **V701** | Husband/Partner’s education |  |  | **0.003** |
|  | *No education* |  16 ((4.2) | 61 (4.4) |  |
|  | *Primary* |  51 (13.3) | 154 (11.2) |  |
|  | *Secondary* |  282 (73.4) | 919 (66.7) |  |
|  | *Higher* |  35 (9.1) | 244 (17.7) |  |
|  | **Empowerment variables** |  |  |  |
| **VV743a** | Respondent decides her health care |  |  | **0.029** |
|  | *No* | 14 (3.5) | 85 (6.0) |  |
|  |  *Yes* | 385 (96.5) | 1326 (94.0) |  |
| **VV743b** | Respondent decides large household purchase  |  |  | **0.019** |
|  | *No* | 45 (11.3) | 99 (7.0) |  |
|  |  *Yes* | 353 (88.7) | 1312 (93.0) |  |
| **VV743d** | Respondent decides on who to visit |  |  | 0.517 |
|  | *No* | 26 (6.5) | 77 (5.5) |  |
|  |  *Yes* | 373 (93.5) | 1334 (94.5) |  |
| **V745a** | Respondent owns a house  |  |  | 0.757 |
|  | *No* |  208 (52.3) | 724 (51.1) |  |
|  |  *Yes* |  190 (47.7) | 693 (48.9) |  |

Note: Mean ±standard deviation is reported for the continuous variables, while binary variables are represented as frequency (%). Reported statistics are population-weighted. Bolded p-values are significant at the 5% level of significance.

The comparative performance for the prediction models is presented in Table 3. With AUC values ranging from 0.704 – 0.758, the four models were able to discriminate between respondents who are victims of IPV and those who are not. RF had the highest AUC value (0.758) and is superior to the other methods in terms of discriminatory power. In terms of specificity, RF also produced the highest value (99.7%), closely followed by GB (99.0%). In terms of balanced accuracy (65.3%), sensitivity (36.7%), and F1 score (46.5%), DT achieved the highest prediction performance. GB outperformed the other methods in terms of precision.

Based on the results of the prediction performances, we observed that DT and RF models had superior performance in most of the considered metrics. However, to construct an interpretable classification model for identifying important risk factors and complex interactions among variables, we favoured DT over RF. Figure 1 shows the rules generated by the DT model.

**Table 3:** Performance measures of the IPV prediction models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  Algorithm | AUC  | Balanced Accuracy (%) | Sensitivity (%) | Specificity (%)  | Precision (%)  | F1 score (%)  |
| RF | **0.758** | 52.6 | 5.5 | **99.7** | 83.3 | 10.4 |
| GB | 0.753 | 58.4 | 17.8 | 99.0 | **84.2** | 29.4 |
| DT | 0.704 | **65.3** | **36.7** | 93.9 | 63.5 | **46.5** |
| LR | 0.746 | 62.2 | 26.7 | 97.8 | 77.4 | 39.7 |

Note: Bolded values indicate the best in performance

Figure 1 shows that the data was first split using the variable ‘fear of the husband/partner’, suggesting that this was the most critical factor in determining a woman’s risk of IPV. Women who feared their husbands/partners were more likely (49%, node 3) to experience IPV relative to 17.0% (node 2) of those who did not. Among women who did not fear their husbands/partners, those who justified wife-beating had a higher risk (57.7%, node 4) of experiencing IPV compared to 14.8% (node 5) of those who did not.

For women who did not fear their husbands/partners and did not justify wife-beating, those whose mothers had a history of IPV were more likely (29.2%, node 8) to experience IPV. In comparison, only 11.0% (node 9) of those with the same characteristics but whose mothers did not have a history of IPV had experienced IPV. Among women who did not fear their husbands/partners, did not justify wife-beating and had mothers who had a history of IPV, those whose husbands/partners took drugs had a higher risk (81.8%, node 15) of experiencing IPV relative to 26.1% (node 14) of those with a similar profile but whose husbands/partners did not take drugs

Among women who did not fear their husbands/partners, those whose husbands/partners drank alcohol had a higher likelihood of experiencing IPV (68.9%, node 7) than 28.4% (node 6) of those whose husbands/partners did not drink alcohol. Again, for women who did not fear their husbands/partners, had husbands/partners who do not drink alcohol, having a mother with a history of IPV increased the risk of IPV (85.7%, node 10). In contrast, only 19.3% (node 11) of women with the same characteristics but whose mothers did not have a history of IPV had experienced IPV.



**Figure 1:** Decision tree classification of IPV

These findings demonstrate informative interaction patterns that profile the characteristics of the IPV victims. In summary, factors associated with IPV were her attitude towards wife-beating, whether her mother was a victim of IPV. We also found that husband characteristics such as drug and alcohol use as well as whether husband/partner instilled fear in the wife increased the risk, with the latter being the most significant factor.

**Discussion**

In this study, we built machine learning models that can efficiently identify the risk of IPV using data from a nationally representative survey in South Africa. As far as we are aware, this is the first study to utilise machine learning to predict vulnerability to IPV in a South African context, as well as using nationally representative data.

Our findings demonstrate that, based on all considered metrics, logistic regression had inferior predictive performances to the other methods. Our study results also showed that the decision trees and random forests models outperformed the other methods. Though the random forest model had a higher specificity, however, since the decision tree model had higher accuracy and sensitivity, it is desirable for prediction. This is because, from a health perspective of IPV, a model with higher sensitivity is more important as it can correctly identify women who are truly vulnerable to IPV. In other words, the consequence of not rescuing a true IPV victim outweighs the consequence of identifying a woman who is not a victim of IPV. Of course, the preferred model could change keeping in mind what might be of ultimate interest for pragmatic interventions.

Our results from the decision tree model show that the following covariates were associated with IPV: fear of the husband or partner, attitudes towards violence, history of abuse, alcohol and drug use were the variables associated with the experience of IPV. Essentially our findings confirm previous studies which have shown a link between IPV and most of these covariates. For instance, multi-country studies and meta-analyses have shown the relationship between substance use, including alcohol and other drugs, and IPV (WHO, 2015; Devries et al. 2015; Kishor & Johnson, 2004; Green et al., 2017) due to its inhibiting and instigating forces (Leonard & Quigley, 2017). In one South African study, approximately 65% of the victims of IPV reported that their husbands/partners were drunk prior to the abuse (O’Connor et al. 2011). As mentioned earlier, there are also studies which have shown that children who have witnessed violence between their parents are more likely to grow up with views about the appropriateness of marital aggression (Islam et al. 2014).

The most striking result to emerge from our analysis is that the fear of the husband or partner was the strongest predictor of IPV amongst South African women. However, this has been the least investigated correlate in the IPV literature in general although some scholars have started speculating that fear might not be an unintentional consequence of violence, but rather something perpetrators recognise, use and play on (Pain, 2014). It has also been opined that in the South African setting, and Africa generally, men are more likely to take advantage of women who are timid and are naturally afraid of them (National Department of Health et al., 2019). Fear of the husband or partner is also common among women with low socio-economic status, marry late or are fearful of divorce. When a man notices this fear, they tend to take advantage of their partners and perpetrate violence because they are confident that their partners can neither leave them nor report to the police authorities (National Department of Health et al., 2019). As plausible as these reasons may be, we also acknowledge the possibility of endogeneity between the fear of a husband or partner and IPV due to reverse causality. For instance, IPV is likely to create a climate of fear where victims live in constant fear of when the next violent episode might occur (Lindgren and Renck 2009). However, our analytical approach did not control for this; hence the results must be interpreted with caution.

Essentially, the splits from the root node to the branches also show the variables that have a more significant effect on IPV, in this case, the fear of the husband/partner was the most significant path. The risk of IPV also varied according to these variables from 17% for women who had non-permissive attitudes towards IPV to 85% for those who had permissive attitudes, had a partner who drinks beer as well as a history of abuse. Thus, by using this machine learning algorithms such as decision tree, researchers can detect the specific combinations of factors that constitute the highest (or lowest) risk for IPV. Such models can also be used to develop interventions by stakeholders such as social workers, policymakers or other interested partners

Regardless of our findings, it is crucial to emphasise that our cross-sectional study cannot guarantee accurate predictions. Hence, our ﬁndings need to be validated using additional studies such as longitudinal or cohort studies. Second, the outcome variable was measured based on self-reported responses; thus, there is the possibility of recall bias on the part of the study respondents. Third, the decision tree model is not robust to bias resulting from other essential risk factors that were not included in the model. Indeed, several studies have identified diverse characteristics that are associated with IPV, including mental health, criminal history etc. It is recommended that future studies might include more factors to improve the predictive power of the model. Our findings should thus be interpreted in light of the explanation above.

**References**

Abramsky, T., Watts, C., Garcia-Moreno, C., Devries, K., Kiss, L. & Ellsberg. M. (2011). What factors are associated with recent intimate partner violence? Findings from the WHO multi-country study on women’s health and domestic violence. *BMC Public Health,* 11 (1), 109.

Abrahams, N., Jewkes, R. & Mathews, S. (2010). Guns and gender-based violence in South Africa. *S Afr Med J*, 100 (9), 586-588.

Abrahams, N., Mathews, S., Martin, L.J., Lombard, C. & Jewkes, R. (2013) Intimate Partner

Femicide in South Africa in 1999 and 2009. *PLoS Med*, 10(4), e1001412. https://doi.org/10.1371/journal.pmed.1001412

Abrahams, N., Mathews, S., Jewkes R., Martin, L.J. & Lombard, L. (2012) Every eight hours: Intimate femicide in South Africa 10 years later! *South African Medical Research C Council Research Brief.* Retrieved from

 http://www.mrc.ac.za/policybriefs/everyeighthours.pdf Accessed 22 February 2013.

Abrahams, N., Jewkes, R., Laubscher, R., & Hoffman, M. ( 2006). Intimate partner violence:

Prevalence and risk factors for men in Cape Town, South Africa. *Violence and Victims*, 21, 247-264. DOI: 10.1891/vivi.21.2.247.

Allen, M., & Devitt, C. (2012). Intimate partner violence and belief systems in Liberia. *Journal of Interpersonal Violence*, 27(17), 3514-3531.

Bisaso, K.R., Anguzu, G.T., Karungi, S.A. & Kiragga, A. (2017). Castelnuovo B. A survey of machine learning applications in HIV clinical research and care. *Comput Biol Med,*  1;91:366-71.

Breiman, L., Friedman, J.H., Olshen, R.A & Stone, C.T. (1984). *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth.

Breiman, L. (2001). Random forests. *Machine Learning*,45(1), 5–32.

Borgoni, R. & Berrington, A. (2013). Evaluating a sequential tree-based procedure for multivariate imputation of complex missing data structures. *Quality & Quantity*, 47(4), 1991–2008.

Chikhungu, L. C., Amos, M., Kandala, N., & Palikadavath, S. (2019). Married Women’s Experience of Domestic Violence in Malawi: New Evidence From a Cluster and Multinomial Logistic Regression Analysis. *Journal of Interpersonal Violence*.

 https://doi.org/10.1177/0886260519851782.

Dalal, K., Wang, S. & Svanstrom, L. (2014). Intimate partner violence against women in Nepal: An analysis through individual, empowerment, family and societal level factors.  *J. Res. Health Sci,* 14(4), 251–257.

Dekel, B. & Andipatin, M. (2016). Abused Women's Understandings of Intimate Partner Violence and the Link to Intimate Femicide. *Qualitative Social Research*, 17(1). Retrieved from research.net/index.php/fqs/article/view/2394/3938.

Devries, K. M., Child, J. C., Bacchus, L. J., Mak, J., Falder, G., Graham, K. & Heise, L. (2014). Intimate partner violence victimization and alcohol consumption in women: A systematic review and meta-analysis. *Addiction*, 109(3), 379-391. https://doi.org/10.1111/add.12393.

Delen, D., Walke,r G. & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med*. 2005;34(2):113–27. DOI: 10.1016/j.artmed.2004.07.002

Eriksson, L., & Mazerolle, P. (2015). A Cycle of Violence? Examining Family-of-Origin Violence, Attitudes, and Intimate Partner Violence Perpetration. *Journal of Interpersonal Violence*, 30(6), 945–964. https://doi.org/10.1177/0886260514539759

Fakir, A.M.S., Anjun, A., Bushra, F. & Nawar, N. (2016). The endogeneity of domestic violence: understanding women empowerment through autonomy. *World Development Perspectives* 2, 34-42. doi.org/10.1016/j.wdp.2016.09.002

Ferreira, D., Oliveira, A. & Freitas, A. (2012). Applying Data Mining Techniques To Improve Diagnosis In Neonatal Jaundice. *BMC Medical Informatics And Decision Making*, 12(1):143. DOI:[10.1186/1472-6947-12-143](https://doi.org/10.1186/1472-6947-12-143)

Friedman, J. (2002). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis,* 38(4), 367–378.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2), 337–407.

Gautam, S. & Jeong, H.S. (2019). Intimate Partner Violence in Relation to Husband Characteristics and Women Empowerment: Evidence from Nepal. *Int J Environ Res Public Health*, 16(5):709. doi: 10.3390/ijerph16050709.

García-Moreno, C., Jansen, H.A., Ellsberg, M., Heis, L., Watt, C. (2005). *WHO multicountry*

*study on women’s health and domestic violence against women: initial results on prevalence, health outcomes and women’s responses*. Geneva: World Health Organization.

Galatzer-Levy, I.R., Ma, S., Statnikov, A., Yehuda, R. & Shalev, A.Y. (2017). Utilisation

of machine learning for prediction of post-traumatic stress: A re-examination of cortisol in the prediction and pathways to non-remitting PTSD.Transl. *Psychiatry*, 7, e1070.

Gibbs, A., Carpenter, B., Crankshaw, T., Hannass-Hancock, J., Smit, J., Tomlinson, M. &

Butler, L. (2017). Prevalence and factors associated with recent intimate partner violence and relationships between disability and depression in post-partum women in one clinic in eThekwini Municipality, South Africa. *PLoS ONE,* 12(7): e0181236. https://doi.org/10.1371/journal.pone.0181236

Greene, M.C., Kane, J.C. & Tol, W.A. (2017). Alcohol use and intimate partner violence among women and their partners in sub-Saharan Africa. *Glob Ment Health (Camb)*, 1 8;4:e13. doi: 10.1017/gmh.2017.9. PMID: 29230309; PMCID: PMC5719482.

Groves, A. K., Moodley, D., McNaughton-Reyes, L., Martin, S. L., Foshee, V., & Maman, S. (2015). Prevalence, rates and correlates of intimate partner violence among South A African women during pregnancy and the postpartum period. *Maternal and child health journal,* 19(3), 487-495. <https://doi.org/10.1007/s10995-014-1528-6>.

Ghosh, D. (2007). Predicting vulnerability of indian women to domestic violence incidents.

 *Research and Practice in Social Sciences*, 3(1), 48-72.

Gordon, C. (2016). Intimate partner violence is everyone’s problem, but how should we approach it in a clinical setting? *South African Medical Journal*, 106 (10), 962-965.

Hosmer, D.W., Lemeshow, S. & Sturdivant, R.X. (2013). *Applied Logistic Regression*. New

 Zealand: Wiley; 2013.

Islam, T.M., Tareque, M.I., Tiedt, A.D. & Hoque, N. (2014). The intergenerational transmission of intimate partner violence in Bangladesh. *Glob Health Action*, 23;7:23591. doi: 10.3402/gha.v7.23591. PMID: 24861340; PMCID: PMC4033320.

Jewkes, R., Nduna, M., Shai, N.J. & Dunkle, K. (2012). Prospective Study Of Rape Perpetration By Young South African Men: Incidence & Risk Factors. *PloS ONE*, 7(5):e38210.

Jewkes, R. (2002). Intimate partner violence: causes and prevention. *The Lancet*,359, e: 1423– 1429.

Johnston, H.B. & Naved, R.T. (2008). Spousal Violence In Bangladesh: A Call For A Public Health Response. *J Health Popul Nutr*, 26(3):366.

Kishor, S. & Johnson, K. (2004). *Profiling Domestic Violence: A Multi-Country Study. Demographic And Health Survey Report.* Calverton MD: ORC Macro, 2004.

Krug, E.G., Mercy, J.A., Dahlberg, L.L. & Zwi, A.B. (2002). The world report on violence and health. *Lancet*. 2002;360(9339):1083–8.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*,521, 436–444.

Lemon, S,C., Roy, J., Clark, M.A., Friedmann, P.D. & Rakowski, W. (2003). Classification

and regression tree analysis in public health: methodological review and comparison with logistic regression. *Soc Behav Med*. 2003;26(3):172–81.

Leonard, K.E. & Quigley, B.M. (2017). Thirty years of research show alcohol to be a cause of intimate partner violence: future research needs to identify who to treat and how to treat them. *Drug and Alcohol Review,* 36, 7–9.

Lindgren, M.S. & Renck, B. (2008). Intimate partner violence and the leaving

process: Interviews with abused women*. International Journal of Qualitative Studies on Health and Well-being*, 3, 113-124. doi: 10.1080/17482620801945805

Mitra, S., Pal, S.K. & Mitra, P. (2002). Data mining in soft computing framework: a survey.

 *IEEE Trans Neural Netw*, 3(1):3–14

National Department of Health (NDoH), Statistics South Africa (Stats SA), South African

Medical Research Council (SAMRC), and ICF. (2019). *South Africa Demographic and Health Survey 2016.* Pretoria, South Africa, and Rockville, Maryland, USA: NDoH, Stats SA, SAMRC, and ICF.

O’Connor, M.J., Tomlinson, M., LeRoux, I.M., Stewart, J., Greco, E. & Rotheram-Borus,

M.J.(2011). Predictors of Alcohol Use Prior to Pregnancy Recognition among Township Women in Cape Town, South Africa. *Soc. Sci. Med*. 2011, 72, 83–90.

Occean, J. R., Thomas, N., Lim, A. C., Lovett, S. M., Michael-Asalu, A., & Salinas-Miranda, A. A. (2020). Prevalence and factors associated with intimate partner violence among women in Haiti: understanding household, individual, partner, and relationship characteristics. *Journal of interpersonal violence*.

 <https://doi.org/10.1177/0886260519898443>.

Pain, R. (2014). Everyday terrorism: Connecting domestic violence and global

terrorism.  *Progress in Human Geography*, *38*(4), 531–550. <https://doi.org/10.1177/0309132513512231>

Petering, R., Um, M. Y., Fard, N. A., Tavabi, N., Kumari, R., & Gilani, S. N. (2018). *Artificial Intelligence to Predict Intimate Partner Violence Perpetration*. In Artificial Intelligence and Social Work, Tambe , M. and Rice E (Eds). Cambridge University Press, Cambridge, pp 195.

Reichel, D. (2017). Determinants of Intimate Partner Violence in Europe: The Role of Socioeconomic Status, Inequality, and Partner Behavior. *Journal of Interpersonal Violence,* 32(12), 1853–1873. <https://doi.org/10.1177/0886260517698951>.

SAS Enterprise Miner 14.2. SAS Institute Inc., Cary, NC.

Shamu, S., Abrahams, N., Temmerman, M., Musekiwa, A. & Zarowsky, C. (2011). A Systematic Review of African Studies on Intimate Partner Violence against Pregnant Women: Prevalence and Risk Factors. *PLoS ONE* 6(3): e17591. https://doi.org/10.1371/journal.pone.0017591

StataCorp. (2015). *Stata Statistical Software*: Release 14. College Station, TX: StataCorp LP.

Sugg, N. (2015). Intimate partner violence: prevalence, health consequences, and intervention. *Medical Clinics*, 99(3), 629-649. <http://dx.doi.org/10.1016/j.mcna.2015.01.012>.

Taylor, R. A., Moore, C. L., Cheung, K. H., & Brandt, C. (2018). Predicting urinary tract infections in the emergency department with machine learning. *PloS one*, 13(3), e0194085. https://doi.org/10.1371/journal.pone.0194085.

Uddin, S., Khan, A., Hossain, M. Moni, MA. (2019). Comparing different supervised machine learning algorithms for disease prediction*. BMC Med Inform Decis Mak* 19, 281. https://doi.org/10.1186/s12911-019-1004-8.

Vetten, L. (2014). *Domestic violence in South Africa*. Institute for Security Studies, Policy Brief 71.

World Health Organization (2015). *Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence*. Geneva: WHO.

Yang, F., Wang, H.Z., Mi, H., Lin, C.D. & Cai, W.W. (2009). Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC Bioinform*. 10, S22.

Yaya, S., Kunnuji, M.O.N. & Bishwajit, G. (2019). Intimate Partner Violence: A Potential Challenge for Women’s Health in Angola. C*hallenges*, 10 (1), 21.

Yilmaz, O. (2018). Female Autonomy, Social Norms and Intimate Partner Violence against

 Women in Turkey. *The Journal of Development Studies,* 54 (8): 1321-1337.

Zembe, Y.Z., Townsend, L., Thorson, A., Silberschmidt, M., Ekstrom, A.M. (2015). Intimate

Partner Violence, Relationship Power Inequity and the Role of Sexual and Social Risk Factors in the Production of Violence among Young Women Who Have Multiple Sexual Partners in a Peri-Urban Setting in South Africa*. PLoS ONE,* 10(11): e0139430. <https://doi.org/10.1371/journal.pone.0139430>.