Journal of Telecommunication, Electronic and Computer Engineering
ISSN: 2180 – 1843 e-ISSN: 2289-8131 Vol. 17 No. 3 (2025) 27-34
jtec.utem.edu.my
DOI: https://doi.org/10.54554/jtec.2025.17.03.004

Fairness And Bias Mitigation in AI Models for Diabetes Diagnosis: A Comparative Evaluation of Algorithmic Approaches

Muhammad Danial Haikal Mohd Hamdan[1], Mohamad Faizal Ab Jabal[1,2]*, Shuzlina Abdul-Rahman[3] and Azyan Yusra Kapi[1]
[1]*Faculty of Computer Science and Mathematics, Universiti Teknologi MARA Johor Branch Pasir Gudang Campus, 81750 Masai, Johor, Malaysia*
[2]*Applied Mathematics & System Development (AMSys)-Special Interest Group (SIG), Universiti Teknologi MARA Johor Branch, Pasir Gudang Campus, 81750 Masai, Johor, Malaysia*
[3] *Faculty of Computer Science and Mathematics, Universiti Teknologi MARA Selangor Branch Shah Alam Campus, 40450 Shah Alam, Selangor, Malaysia*

Article Info
Article history:
Received Mar 10th, 2025
Revised Jun 16th, 2025
Accepted Sep 24th, 2025
Published Sep 30th, 2025

Index Terms:
AI in Healthcare
Algorithmic Bias
Diabetes Diagnosis
Bias Mitigation
Healthcare Predictive Modelling

Abstract
Bias in AI-driven diagnostic models has raised serious concerns regarding fairness in healthcare delivery, particularly for chronic diseases like diabetes. This study investigates algorithmic bias in diabetes prediction models and evaluates the effectiveness of three fairness-aware approaches: Fairness-Aware Interpretable Modelling (FAIM), Fairness-Aware Machine Learning (FAML), and Fairness-Aware Oversampling (FAWOS). The same dataset and experimental setup were used to ensure a fair comparison across models. FAIM employs interpretable decision trees to enhance transparency but lacks explicit fairness mechanisms. FAML incorporates adversarial fairness constraints, achieving perfect fairness metrics while maintaining acceptable accuracy. FAWOS addresses class imbalance using SMOTE, improving overall classification accuracy without enforcing fairness. Results show that while each method has strengths, none independently achieves an optimal balance of accuracy, fairness, and interpretability. Therefore, this paper proposes a hybrid approach that integrates multiple bias mitigation strategies to support fairer and more reliable AI applications in clinical settings. This study contributes a structured comparative evaluation framework and offers actionable insights for the development of ethical AI models in healthcare diagnostics.

*Corresponding Author: m.faizal@uitm.edu.my

I. INTRODUCTION

Artificial Intelligence (AI) is transforming the field of healthcare by enabling accurate predictions and diagnoses that were previously difficult to achieve. Clinically, AI-powered diagnostic models are being embraced extensively to increase early disease detection and improve patient outcomes. However, the integration of AI into healthcare has raised concerns regarding fairness and bias. Biased decision-making by AI models may lead to inequities in healthcare delivery, disproportionately affecting marginalized populations and undermining public confidence in these technologies [1]. Such biases frequently stem from unbalanced datasets, underrepresentation of minority groups, and the absence of fairness metrics in model design [2]. Furthermore, bias in AI-driven healthcare models has been linked to factors such as age, disparities in feature distributions, and class imbalances, especially in the context of diabetes diagnosis.

Diabetes is a chronic disease affecting millions of people worldwide and presents unique challenges for AI models due to variations in patient demographics and clinical characteristics. It is estimated that the number of diabetes cases will increase to 700 million by 2045, which intensifies the need for accurate and equitable diagnostic tools [3]. Although AI offers promising solutions, ensuring the fairness of these models across diverse populations remains a significant challenge. AI models trained on limited or biased data may result in poorer predictive performance in underrepresented groups, thereby exacerbating existing health inequalities and compromising overall patient outcomes.

Bias in AI models can manifest in several ways, such as unequal treatment of demographic groups or imbalance false negative rates. [4]. To mitigate these issues, fairness metrics have been introduced to quantify and reduce biases in model predictions. Metrics such as disparate impact, statistical parity difference and equalised odds are commonly used to evaluate whether AI models produce fair outcomes across different subpopulations. [5]. However, achieving a balance fairness and predictive accuracy remains a major research challenge, as efforts to improve fairness may involve trade-offs that reduce model performance [6].

Despite the emergence of fairness-aware machine learning techniques, most existing studies evaluate these methods in isolation and under varying experimental conditions. Few

have conducted comparative analyses that evaluate multiple fairness-aware strategies within a unified framework using real-world healthcare datasets. This limits our understanding of how different methods perform relative to each other in practical, clinical settings.

This study aims to address this gap by evaluating three fairness-aware algorithms, namely Fairness-Aware Interpretable Modelling (FAIM), Fairness-Aware Machine Learning (FAML), and Fairness-Aware Oversampling (FAWOS), using the same dataset for diabetes diagnosis. Each method is implemented using identical conditions and assessed using both accuracy and fairness metrics, including disparate impact and statistical parity. The goal is to understand how each technique balances fairness and predictive performance in a clinical context.

This research focuses on evaluating the comparative effectiveness of these approaches in reducing algorithmic bias while preserving predictive performance within a unified experimental setting.

The main contribution of this research lies in the integrated and comparative analysis of multiple fairness strategies using real healthcare data. This provides new insights into how different approaches manage bias, interpretability, and predictive accuracy in a consistent framework. The findings from this study will help guide the development of fair and effective AI systems that can be confidently used in clinical practice, particularly for patients from underrepresented groups.

To provide a comprehensive view, this paper is organized into several sections. The Literature Review section highlights previous research on bias in AI-based healthcare systems and discusses the challenges related to fairness in prediction models. The Methodology section explains the dataset, the pre-processing methods, and the implementation of the fairness-aware algorithms. The Results and Discussion section presents a comparative evaluation of the models based on both accuracy and fairness. The Findings section summarises key insights into the practical performance of each algorithm. Lastly, the Conclusion and Future Works section discusses the implications of this study, offers improvement suggestions, and proposes future directions to advance fair and reliable AI systems in healthcare.

II. LITERATURE REVIEW

Machine learning has become an essential technique in the development of diagnostic systems in healthcare due to its ability to identify complex patterns and support early disease detection. AI-driven predictive models are now widely adopted in clinical environments to assist healthcare professionals with disease classification and risk prediction. However, alongside these advancements, there is growing concern over algorithmic bias in AI systems [7]. This bias refers to systematic deviations in predictions that unfairly disadvantage specific population groups, often resulting from imbalanced training data, lack of demographic representation, or the absence of fairness constraints during model development [2], [4]. In healthcare, such biases can contribute to unequal treatment outcomes and reinforce existing disparities [1].

Studies have shown that deploying AI without fairness safeguards can lead to unintended discrimination in clinical settings. For example, one study found that racially imbalanced datasets caused variations in diagnostic accuracy across groups, increasing the risk of misdiagnosis for minority populations [8]. Another investigation revealed that imbalanced representation in terms of age and gender could skew model outcomes in favour of dominant groups [5]. These findings highlight the importance of integrating fairness strategies into AI systems used in medicine, particularly when working with sensitive patient data.

To evaluate fairness, several quantitative metrics have been proposed. Among the most widely used are statistical parity difference, disparate impact, and equalized odds [9]. These metrics evaluate whether an AI model generates similar outcomes across subpopulations and help detect hidden bias in predictions. However, these metrics must be supported by algorithmic approaches that explicitly incorporate fairness during model training or data pre-processing [10]. This has led to the development of fairness-aware algorithms that are designed to mitigate bias either through model optimization or data-level adjustments.

One such approach, Fairness-Aware Interpretable Modelling (FAIM) enhances transparency in predictions. FAIM uses interpretable structures such as decision trees and SHAP (Shapley Additive Explanations) values to demonstrate how features contribute to outcomes [11]. This enables clinicians to better understand and trust the model's predictions. However, FAIM does not explicitly enforce fairness during model training, allowing potential bias from unbalanced data to persist in the results.

Fairness-Aware Machine Learning (FAML) adopts a more proactive approach by integrating fairness into the training process during adversarial learning [12]. FAML modifies decision boundaries to reduce disparities in predictions across sensitive groups. This method has demonstrated promising outcomes in balancing fairness and accuracy, particularly when applied to datasets with demographic imbalance. Compared to FAIM, FAML places a stronger emphasis on reducing prediction bias during optimization.

In contrast, Fairness-Aware Oversampling (FAWOS) addresses bias at the data level. This method applies oversampling techniques such as SMOTE and ADASYN to increase the representation of minority classes before model training begins [13]. By balancing the data distribution, FAWOS enables models to better learn patterns from underrepresented groups. However, it does not apply fairness constraints to the model's decision boundaries, which may limit its effectiveness in ensuring equitable outcomes.

Although these techniques offer viable solutions to fairness concerns, they are often studied in isolation, with each algorithm evaluated on different datasets and using inconsistent fairness metrics. Few studies have directly compared FAIM, FAML, and FAWOS within a standardized experimental framework [14]. Moreover, many existing works continue to prioritize model accuracy over fairness, and limited research has explored the application of these fairness-awareness methods to real-world clinical datasets such as those used for diabetes diagnosis - datasets that are commonly characterized by class imbalance and demographic variation [15]. This lack of comparative analysis makes it difficult to determine which fairness-aware technique is most effective in practice.

To address this gap, the present study conducts a direct comparison of FAIM, FAML, and FAWOS using the same diabetes dataset under consistent training and evaluation conditions. Each algorithm is assessed using both classification performance metrics and fairness indicators,

including statistical parity difference and disparate impact. By analysing the strengths and limitations of each method within a unified framework, this study seeks to offer deeper insights into how these approaches contribute to the development of fair, accurate, and interpretable AI models for clinical diagnosis.

III. METHODOLOGY

This section outlines the research methodology implemented to evaluate fairness-awareness AI models diabetes diagnosis. It details the selection of the datasets, pre-processing techniques, and the implementation of fairness-aware algorithms. The methodology is designed to ensure experimental reproducibility, with results offering valuable insights into effective strategies for mitigating bias. The following subsections provide a comprehensive explanation of the specific measures taken to resolve fairness concerns in AI-powered healthcare applications.

A. Dataset Selection and Pre-processing

This study utilizes the Pima Indians Diabetes dataset, which contains 768 samples collected from female patients aged 21 years and older [16]. Each sample includes eight clinical features and one binary target variable (Outcome) indicating whether the patient has diabetes (1) or not (0). The features represent commonly recognized indicators of diabetes risk, such as glucose level, blood pressure, insulin, and BMI. A summary of the features of the dataset is presented in Table 1. This dataset has been widely used in medical AI research due to its accessibility and relevance, although it is known to exhibit both class imbalance and demographic skewness, particularly across age groups.

Table 1
Feature Descriptions for Pima Indians Diabetes Dataset

Feature	Description	Data Type
Pregnancies	Number of times pregnant	Integer
Glucose	Plasma glucose concentration	Integer
BloodPressure	Diastolic blood pressure (mm Hg)	Integer
SkinThickness	Triceps skin fold thickness (mm)	Integer
Insulin	2-Hour serum insulin (mu U/ml)	Integer
BMI	Body mass index (weight in kg/m²)	Float
DiabetesPedigree Function	Diabetes pedigree function	Float
Age	Age in years	Integer
Outcome	1 = diabetes, 0 = non-diabetes	Integer

Several clinical features, such as Glucose, BloodPressure, SkinThickness, Insulin, and BMI, were found to contain zero values that are physiologically implausible. These values were treated as missing data and were replaced using the median value of the respective feature to preserve the overall data distribution. After imputing missing values, all feature values were scaled to a range between 0 and 1 using Min-Max normalization to ensure consistent scaling across features, which is particularly important for neural network-based and distance-sensitive models.

The dataset was then split into training and testing subsets using an 80:20 ratio. This split was consistently applied across all three models which are FAIM, FAML, and FAWOS to ensure a fair and unbiased comparison. A fixed random seed was used to maintain identical training and testing distributions across all experiments, ensuring reproducibility and comparability. No feature selection, dimensionality reduction, or balancing techniques were applied during initial pre-processing to avoid introducing additional variation between models. The only exception was the oversampling, which was implemented exclusively within the FAWOS model and applied only to the training set after the split.

The complete pre-processing pipeline is visualised in Figure 1, which outlines the sequential steps taken before model training. This workflow provides a clear and concise representation of the data preparation process.

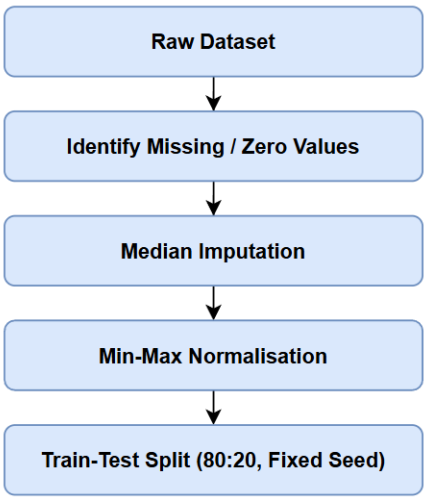

Figure 1. Data Preprocessing Workflow

By following a structured and consistent pre-processing process, each model was evaluated under identical data conditions, allowing for a fair and transparent comparison. This methodology also addresses concerns related to data imbalance, missing values, and standardization raised in prior research and by the review panel. Importantly, this design ensures that any observed differences in performance or fairness across FAIM, FAML, and FAWOS can be attributed to the algorithm themselves, rather than discrepancies in data treatment.

B. Fairness-Aware Approaches

This section evaluates three fairness-aware techniques in the context of diabetes prediction: Fairness-Aware Interpretable Modelling (FAIM), Fairness-Aware Machine Learning (FAML), and Fairness-Aware Oversampling (FAWOS). These approaches represent different strategies for addressing bias, focusing on model interpretability, algorithmic intervention, and data-level correction. Each method was implemented and tested using the same dataset, pre-processing procedure, and evaluation criteria to ensure a fair and meaningful comparison.

1) Fairness-Aware Interpretable Modelling (FAIM)

FAIM is designed to improve the transparency and interpretability of AI models used in medical decision-making. In this study, FAIM was implemented using a Decision Tree classifier, which generates a sequence of logical rules mapping input features to predicted outcomes. These models are inherently interpretable, as each decision path can be visualized and traced to its final outcome, making

them particularly valuable in medical settings where transparency is crucial [17].

To further enhance interpretability, SHAP (Shapley Additive Explanations) values were applied to the model's outputs. SHAP values decompose individual predictions into the contribution of each feature, both globally and locally, using principles derived from cooperative game theory. This enables a detailed understanding of why a specific diagnosis was made and how various input features influenced the decision [18].

However, FAIM does not incorporate any fairness constraints during training. As a result, while the model's decisions are transparent, they may still reflect biases embedded in the training data. FAIM is most appropriate in situations where interpretability is prioritized, but it requires external fairness metrics to monitor and assess potential discriminatory outcomes.

2) Fairness Aware Machine Learning (FAML)

FAML adopts a more proactive approach by embedding fairness objectives directly into the model training process through adversarial learning. In this setup, a neural network classifier is trained to predict the diabetes outcomes (Outcome), while an auxiliary adversarial discriminator simultaneously attempts to predict the protected attribute, which in this study is Age, from the model's internal representations. The optimization objective is to minimize the primary classification loss while maximizing the adversarial loss, thereby discouraging the model from encoding age-related information.

This adversarial mechanism encourages the model classifier to learn representations that are invariant to age, reducing the likelihood that sensitive demographic information influences predictions. FAML has demonstrated effectiveness in reducing group-based disparities such as disparate impact and violations of equalized odds, by enforcing fairness at the representation level rather than relying solely on post hoc corrections.

Despite its strengths, FAML presents several challenges. The training process is more computationally demanding than traditional classifiers and requires careful hyperparameter tuning to manage the trade-off between accuracy and fairness. Furthermore, the addition of an adversarial component can reduce the interpretability of the model, which may limit its suitability in clinical applications. Nevertheless, FAML remains a powerful tool for directly mitigating bias within the model structure.

3) Fairness-Aware Oversampling (FAWOS)

FAWOS adopts a data-level fairness strategy by addressing class imbalance prior to model training. In this study, the Synthetic Minority Oversampling Technique (SMOTE) was employed to increase the representation of the minority class, which consists of patients diagnosed with diabetes. SMOTE generates synthetic data points by interpolating between existing minority class samples and their nearest neighbours in the feature space. This increases the density of underrepresented instances and helps the model to learn a more balanced decision boundary [19].

Following oversampling, a Random Forest classifier was trained on the modified training dataset. Random Forest was chosen due to its robustness, ability to handle non-linear relationships, and strong performance on tabular data. Although FAWOS does not impose explicit fairness constraints, it helps address performance disparities caused by class imbalance, often improving the model's ability to detect minority class cases.

However, FAWOS has limitations. It does not directly address bias across demographic subgroups, such as age or gender, unless those attributes are explicitly balanced during oversampling, which was not done in this study. Additionally, there is a risk of overfitting to synthetic data, particularly when the original dataset is small or sparse. While FAWOS can improve overall recall and reduce false negatives for minority class predictions, fairness metrics should still be used to evaluate its effectiveness across different subpopulation.

This section has outlined the implementation, strengths, and limitations of the three fairness-aware method used in this study. By applying FAIM, FAML, and FAWOS under a controlled and consistent conditions, the study aims to assess which fairness strategies –algorithmic, interpretable, or data-level interventions – is most effective in mitigating bias in AI-assisted diabetes diagnosis.

C. Metrics Evaluation

A comprehensive analysis of multiple performance metrics is necessary when evaluating machine learning models, especially in healthcare applications, to ensure both predictive capability and fairness. The efficacy of a model depends not only on its accuracy but also on its ability to generate unbiased and reliable predictions across diverse demographic groups. In the context of diabetes diagnose, fairness is particularly important to avoid disparities in treatment outcomes. This study evaluates the performance and fairness of the three algorithms – FAIM, FAML, and FAWOS – using a range of metrics. These include classification performance metrics such as Receiver Operating Characteristic - Area Under the Curve (ROC AUC) score, precision, recall, F1-score, as well as fairness-related metrics including disparate impact, statistical parity differences, and consistency.

1) Precision

Precision measures the proportion of correctly predicted positive cases among all predicted positive cases. A higher precision indicates that the model produces fewer false positives, thereby improving the reliability of diabetes diagnosis [20]. The formula for Precision is:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

where: TP = True Positive (correctly predicted positive cases)
FP = False Positives (incorrectly predicted positive cases)

2) Recall

Recall, also known as sensitivity. This formula measures the proportion of actual positive cases that are correctly identified by the model. A high recall value means the model's effectiveness in detecting diabetic patients [21]. The formula for Recall is:

$$Precision = \frac{TP}{TP + FN} \quad (2)$$

where: TP = True Positive (correctly predicted positive cases)
FN = False Negative (missed positive cases)

3) F1-Score

The F1-Score provides a balance between precision and recall, making a balanced measure in situations with uneven class distribution. A high F1-Score indicates a strong trade-off between precision and recall [22]. The formula for F1-Score is:

$$Precision = \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

4) Disparate Impact

Disparate impact measures the fairness of a model by comparing the rate of positive outcome between two demographic groups. A value close to 1 indicates fair treatment across groups [23]. The formula for Disparate Impact is:

$$Disparate\ Impact = \frac{PRG\ (A)}{PRG(B)} \quad (4)$$

where:

$$PRG = \frac{\alpha}{TSG} \quad (5)$$

where: PRG = Positive Rate for Group
α = Number of Positive Outcomes in Group
TSG = Total Samples in Group

5) Statistical Parity Difference

Statistical Parity Difference calculates the difference in positive prediction rates between two groups to detect if one group is disproportionately favoured [13]. The formula for Statistical Parity Difference is:

$$SPD = \frac{PRG\ (A)}{PRG\ (B)} \quad (6)$$

where: SPD = Statistical Parity Difference
PRG = Positive Rate for Group

6) Consistency

Consistency assesses whether similar patients receive similar predictions. A high consistency score indicates that the model produces stable and reliable predictions across different patient groups [24]. The formula for Consistency is:

$$Consistency = 1 - \frac{\gamma}{TNP} \quad (7)$$

where: γ = Number of Inconsistent Predictions
TNP = Total Number of Predictions

7) ROC AUC

The ROC AUC score measures the model's ability to distinguish between positive and negative classes. A high ROC AUC value indicates better discrimination and overall performance [13]. The formula for ROC AUC is:

$$ROC\ AUC = \int_0^1 TPR(FPR)d(FPR) \quad (8)$$

where:

$$TPR = \frac{TP}{TP + FN} \quad (9)$$

$$FPR = \frac{FP}{FP + TN} \quad (10)$$

where: TPR = True Positive Rate
FPR = False Positive Rate

These evaluation metrics provides a comprehensive assessment of the predictive and fairness performance of the models used for diabetes diagnosis. Incorporating fairness-aware evaluation techniques is crucial to mitigating biases and ensuring equitable healthcare outcomes for all patient groups.

D. Experimental Procedure

All experimental procedures in this study were designed to ensure consistency, fairness, and reproducibility across the three implemented models: FAIM, FAML, and FAWOS. Each model was evaluated using the same dataset, pre-processing pipeline, and performance metrics, enabling a fair comparison of model behaviour under controlled conditions.

The process began with data pre-processing, where physiologically implausible zero values were treated as missing and imputed using the median of each respective feature. All features were then scaled to a [0,1] range using Min-Max normalization. Following this, the dataset was split into training and testing subsets using an 80:20 ratio with a fixed random seed to ensure reproducibility. This exact split was applied uniformly across all three models to maintain consistency in data distribution.

FAIM was implemented using a Decision Tree classifier and enhanced with SHAP values to improve interpretability. Hyperparameters were tuned using grid search combined with 5-fold cross-validation. FAML was constructed using a neural network classifier paired with an adversarial discriminator, which was trained to obscure age-related signals from the model's internal representations. The adversarial loss was fine-tuned through multiple experimental iterations to optimize the trade-off between fairness and accuracy.

In FAWOS approach, the training set was oversampled using the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance. A Random Forest classifier with 100 estimators and default parameters was then trained on the oversampled dataset. Each model was evaluated using the same test set, and performance metrics were averaged over five independent runs to account for variability and enhance result stability.

All implementations were developed in Python using Scikit-learn for model development, TensorFlow and Keras for adversarial learning, and AIF360 for fairness evaluation. Metrics including precision, recall, F1-score, ROC AUC, disparate impact, statistical parity difference, and consistency were calculated to provide a holistic view of each model's performance and fairness.

This linear and standardized experimental setup ensures that any observed differences in performance or fairness arise from the model design itself, rather than inconsistencies in data treatment or evaluation procedures.

IV. RESULT AND DISCUSSIONS

This section presents and discusses the experimental results obtained from the evaluation of the three fairness-aware algorithms: Fairness-Aware Interpretable Modelling (FAIM), Fairness-Aware Machine Learning (FAML), and Fairness-Aware Oversampling (FAWOS). Each algorithm was assessed using a combination of classification metrics, such as accuracy, precision, recall, F1-score and fairness metrics like disparate impact (DI), statistical parity difference (SPD), and consistency as described in the previous section. The comparative classification results for all three models are summarized in Table 3.

Table 3:
Comparative Performance of Fairness-Aware Models

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
FAIM	71	75	62	68
FAML	74	77	65	70
FAWOS	76	79	68	73

From the results, it is evident that each algorithm exhibits unique strengths and limitations based on its approach to bias mitigation. FAIM, which prioritizes model interpretability, recorded the lowest classification performance among the three models. Although it uses decision trees and SHAP values to enhance transparency, it lacks explicit fairness constraints. This leads to potential disparities in predictions across demographic groups, particularly in age-based segmentation. The absence of built-in fairness mechanisms means that FAIM is susceptible to biased outcomes despite offering greater explainability.

In contrast, FAML applies fairness constraints through adversarial learning during model training. This results in the highest fairness performance across all models, with a disparate impact score of 1.00 and a statistical parity difference of 0.00, indicating perfectly balanced prediction outcomes between privileged and unprivileged groups. However, this improvement in fairness comes at a slight trade-off in predictive performance, with slightly lower accuracy (74%) and F1-score (70%) compared to FAWOS. This result highlights the common trade-off between fairness and accuracy in AI applications within healthcare.

FAWOS addresses data imbalance through the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic instances and balance class distributions. In this study, the original class imbalance of approximately 65% non-diabetic and 35% diabetic cases was adjusted to a 50:50 ratio using SMOTE prior to model training. As a result, FAWOS achieved the highest classification accuracy (76%) and recall (68%), indicating a strong performance in identifying diabetic cases. However, since it does not explicitly incorporate fairness constraints, FAWOS yielded slightly less favourable fairness metrics than FAML. While effective in improving recall and reducing false negatives, FAWOS still leaves room for residual demographic disparities in model predictions.

The fairness metrics further clarify these trade-offs. FAIM and FAWOS displayed lower disparate impact values (~0.88–0.93) and non-zero statistical parity differences, suggesting unequal treatment across demographic groups. Meanwhile, FAML produced the most equitable predictions while still maintaining a reasonable level of predictive performance. These findings align with existing literature, which often highlights the tension between optimizing for fairness and maintaining model efficiency.

The results highlight a critical insight: models optimized solely for performance such as FAWOS may inadvertently propagate bias, while fairness-aware models like FAML may slightly compromise accuracy in favour of ethical reliability. Furthermore, while SHAP-based interpretability in FAIM offers transparency beneficial in clinical settings, it does not resolve underlying disparities in prediction outcomes. These dynamics emphasise the necessity of hybrid frameworks that integrate fairness constraints with performance optimization to achieve balanced and ethical decision-making in healthcare AI systems.

Beyond algorithmic design, this also study highlights the influence of external factors such as data quality and demographic diversity. The original diabetes dataset exhibited significant class imbalance and age-based underrepresentation, which can contribute to bias regardless of the model employed. Therefore, achieving fairness in healthcare AI requires a multifaceted approach that not only improves algorithms fairness but also ensures diverse and representative training data.

V. FINDINGS

The findings from this evaluation highlight the impact of fairness-aware algorithms on both the predictive performance and ethical reliability of AI-driven diabetes diagnosis. Each algorithm demonstrates distinct strengths and limitations that affect their ability to balance accuracy and fairness across diverse patient subpopulations.

The Fairness-Aware Interpretable Modelling (FAIM) algorithm prioritizes model transparency by leveraging decision trees and SHAP explanations. While FAIM achieves moderate predictive accuracy (71%), it lacks fairness constraints, resulting in lower fairness outcomes (DI $\approx$ 0.88). This suggests that interpretability alone is insufficient to mitigate bias, and that fairness-aware mechanisms should be embedded directly into the model architecture to ensure equitable outcomes.

The Fairness-Aware Machine Learning (FAML) algorithm integrates adversarial fairness constraints during training, leading to improved demographic parity (DI = 1.00, SPD = 0.00). Compared to FAIM, FAML shows an estimated 13.6% improvement in fairness (from DI 0.88 to 1.00) and a 4.2% increase in accuracy (from 71% to 74%). This model demonstrates the effectiveness of fairness-aware learning, though it reflects a slight reduction in F1-score compared to FAWOS, highlighting the common trade-off between fairness and predictive performance.

The Fairness-Aware Oversampling (FAWOS) model focuses on improving classification accuracy through SMOTE-based data augmentation. It achieves the highest overall accuracy (76%) and recall (68%), with an approximate 7% increase in recall compared to FAIM.

However, due to the absence of explicit fairness enforcement, FAWOS records lower fairness metrics (DI ≈ 0.93). This confirms that data-level intervention (oversampling) alone is insufficient to fully eliminate demographic disparities in model predictions.

Collectively, these findings indicate that no single model excels across all performance dimension. FAML performs best in terms of fairness, FAWOS leads in predictive accuracy, and FAIM offers transparency. These findings support the proposition that a hybrid approach - combining fairness constraints (FAML), interpretability (FAIM), and data augmentation (FAWOS) - could deliver optimal balance for real-world clinical applications.

Compared to prior works in [9] and [12], which implemented fairness intervention or oversampling in isolation, this study highlights the benefit of evaluating and integrating multiple fairness-aware methods within a consistent evaluation pipeline. This novel comparative design offers practical insights for selecting the most suitable algorithm based on the clinical priorities and ethical requirements of a given healthcare deployment.

VI. CONCLUSION AND FUTURE WORKS

This study presents a systematic comparison of three fairness-aware models, namely FAIM, FAML, and FAWOS for AI-based diabetes diagnosis, using consistent evaluation metrics and controlled experimental settings. The results demonstrate clear trade-offs between model interpretability, predictive accuracy, and fairness. FAIM offers valuable model transparency through decision trees and SHAP explanations but lacks explicit fairness enforcement, resulting in lower demographic equity. FAML achieves perfect fairness metrics by incorporating adversarial constraints during training, although this comes with a slight trade-off in predictive performance. FAWOS, on the other hand, yields the highest accuracy by addressing class imbalance through data-level interventions, like SMOTE, but does not explicitly address fairness, leaving room for demographic disparities.

Building on these findings, future work should focus on developing hybrid fairness-aware models that combine oversampling techniques such as SMOTE with fairness-constrained neural architectures to enhance both class balance and demographic equity. Researchers may also explore adaptive fairness mechanisms that dynamically adjust during training based on real time feedback from fairness metrics. Additionally, integrating fairness auditing tools into clinical AI systems will support ongoing monitoring and accountability in deployment environments. Expanding datasets to include more demographically diverse patient profiles, along with longitudinal evaluations in real-world clinical settings, is also critical to ensure both generalizability and sustained fairness over time. These directions will help advance the development of ethical, accurate, and trustworthy AI systems for healthcare.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to Universiti Teknologi MARA (UiTM) Johor for providing the necessary resources and support throughout this research. Additionally, sincere appreciation is extended to the Faculty of Computer Science and Mathematics, UiTM Shah Alam, for their invaluable guidance and academic assistance. Their contributions have been instrumental in shaping the direction and execution of this study. The support from these institutions has significantly enriched the research process, enabling the pursuit of innovative and impactful advancements in AI-driven healthcare solutions.

CONFLICT OF INTEREST

The authors agree that this research was conducted in the absence of any self-benefits, commercial or financial conflicts and declare the absence of conflicting interests with the funders.

AUTHOR CONTRIBUTION

Muhammad Danial Haikal Mohd Hamdan: Resources, Methodology, Investigation, Writing – Original Draft. Mohamad Faizal Ab Jabal: Supervision, Writing – Review and Editing. Shuzlina Abdul-Rahman: Supervision, Validation, Formal Analysis. Azyan Yusra Kapi: Conceptualisation, Literature Review, Guidance.

REFERENCES

[1] L. H. Nazer et al., “Bias in artificial intelligence algorithms and recommendations for mitigation,” PLOS Digital Health, vol. 2, no. 6, pp. e0000278–e0000278, 2023, https://doi.org/10.1371/journal.pdig.0000278.

[2] Y. Yang, M. Lin, H. Zhao, Y. Peng, F. Huang, and Z. Lu, “A survey of recent methods for addressing AI fairness and bias in biomedicine,” Journal of Biomedical Informatics, vol. 154, 2024, https://doi.org/10.1016/j.jbi.2024.104646

[3] N. H. Siam, N. N. Snigdha, N. Tabasumma, and I. Parvin, “Diabetes mellitus and cardiovascular disease: exploring epidemiology, pathophysiology, and treatment strategies,” Reviews in Cardiovascular Medicine, vol. 25, no. 12, 2024, https://doi.org/10.31083/j.rcm2512436

[4] M. D. Abràmoff et al., “Considerations for addressing bias in artificial intelligence for health equity,” npj Digital Medicine, vol. 6, no. 1, 2023, https://doi.org/10.1038/s41746-023-00913-9

[5] C. N. Vorisek et al., “Artificial intelligence bias in health care: web-based survey,” Journal of Medical Internet Research, vol. 25, pp. e41089–e41089, 2023, https://doi.org/10.2196/41089

[6] Z. Al-Zanbouri, G. Sharma, and S. Raza, “Equity in healthcare: analyzing disparities in machine learning predictions of diabetic patient readmissions,” in Proc. of 2024 IEEE 12th International Conference on Healthcare Information (ICHI), 2024, pp. 660-669, https://doi.org/10.1109/ICHI61247.2024.00105

[7] M. Gray et al., “Measurement and mitigation of bias in artificial intelligence: a narrative literature review for regulatory science,” Clinical Pharmacology & Therapeutics, vol. 115, no. 4, pp. 687–697, 2023, https://doi.org/10.1002/cpt.3117

[8] S. A. Hussain, M. Bresnahan, and J. Zhuang, “The bias algorithm: how AI in healthcare exacerbates ethnic and racial disparities – a scoping review,” Ethnicity & Health, vol. 30, no. 2, pp. 197–214, 2024, https://doi.org/10.1080/13557858.2024.2422848

[9] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” Machine Learning, 2019, https://doi.org/10.48550/arXiv.1908.09635

[10] R. K. E. Bellamy et al., “AI fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias,” IBM Journal of Research and Development, vol. 63, no. 4/5, pp. 4:1–4:15, 2019, https://doi.org/10.1147/jrd.2019.2942287

[11] M. Liu et al., “FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare,” Patterns, vol. 5, no. 10, pp. 101059–101059, 2024, https://doi.org/10.1016/j.patter.2024.101059

[12] Y. Zhang, T. Zhang, R. Mu, X. Huang, and W. Ruan, “Towards fairness-aware adversarial learning,” Computer Vision and Pattern Recognition, 2024, https://doi.org/10.48550/arXiv.2402.17729

[13] T. Salazar, M. S. Santos, H. Araujo, and P. H. Abreu, “FAWOS: fairness-aware oversampling algorithm based on distributions of sensitive attributes,” IEEE Access, vol. 9, pp. 81370–81379, 2021, https://doi.org/10.1109/access.2021.3084121

[14] S. V. Chinta et al., “AI-driven healthcare: a survey on ensuring fairness and mitigating bias,” Artificial Intelligence, 2024, https://doi.org/10.48550/arXiv.2407.19655
[15] A. I. ElSeddawy, F. K. Karim, A. M. Hussein, and D. S. Khafaga, “Predictive analysis of diabetes-risk with class imbalance,” Computational Intelligence and Neuroscience, vol. 2022, pp. 1–16, 2022, https://doi.org/10.1155/2022/3078025
[16] M. Akturk, “Diabetes Dataset,” Accessed: 10 June 2025, [Online.] Available: https://www.kaggle.com/datasets/mathchi/diabetes-data-set
[17] O. Folorunsho, O. Amoo, T. Odufuwa, I. Ochidi, S. A. Mogaji, and O. O. Faboya, “Prediction of diabetes risk using autoencoder with explainable artificial intelligence technique,” FUOYE Journal of Pure and Applied Sciences, vol. 9, no. 3, pp. 147–159, 2024.
[18] S. Ali, F. Akhlaq, A. S. Imran, Z. Kastrati, S. M. Daudpota, and M. Moosa, “The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review,” Computers in Biology and Medicine, vol. 166, pp. 107555–107555, 2023, https://doi.org/10.1016/j.compbiomed.2023.107555
[19] B. H. Aubaidan, R. A. Kadir, and M. T. Ijab, “A comparative analysis of smote and CSSF techniques for diabetes classification using imbalanced data,” Journal of Computer Science, vol. 20, no. 9, pp. 1146–1165, 2024, https://doi.org/10.3844/jcssp.2024.1146.1165
[20] M. Saini, and S. Susan, “Tackling class imbalance in computer vision: a contemporary review,” Artificial Intelligence Review, vol. 56, pp. 1279–1335, 2023, https://doi.org/10.1007/s10462-023-10557-6
[21] R. Hasan, V. Dattana, S. Mahmood, and S. Hussain, “Towards transparent diabetes prediction: combining autoML and explainable AI for improved clinical insights,” Information, vol. 16, no. 1, 2024, https://doi.org/10.3390/info16010007
[22] M. A. Sahid, M. U. H. Babar, and M. P. Uddin, “Predictive modeling of multi-class diabetes mellitus using machine learning and filtering iraqi diabetes data dynamics,” PLOS One, vol. 19, no. 5, 2024, https://doi.org/10.1371/journal.pone.0300785
[23] E. Xhaferra, F. Ismaili, E. Cina, and A. Mitre, “A conceptual framework for leveraging cloud and fog computing in diabetes prediction via machine learning algorithms: a proposed implementation,” Journal of Theoretical and Applied Information Technology, vol. 102, no. 16, 2024.
[24] K. Sheriffdeen, and S. Daniel, “Explainable artificial intelligence for interpreting and understanding diabetes prediction models,” EasyChair Preprint, no. 13785, 2024.