LFGA–ADENN Hybrid Feature Extraction and Oversampling for Type 2 Diabetes Mellitus Detection in Mizoram

Vanlalawmpuia R[1], Lalhmingliana[*2], Nachimuthu Senthil Kumar[3], Brindha Senthil Kumar[4], Freda Lalrohlui[5], John Zohmingthanga[6], Vanlalhruaii[7].

[1]Department of Computer Engineering, Mizoram University, Aizawl, Mizoram 796004, India
[2]Department of Information Technology, Mizoram University, Aizawl, Mizoram 796004, India
[3, 5]Department of Biotechnology, Mizoram University, Aizawl, Mizoram 796004, India
[4]Sri Eshwar College of Engineering, Vadasithur Road, Vadasithur-641202, India
[6]Zoram Medical College, Falkawn, Mizoram - 796 005, India
[7]Department of Medicine, Zoram Medical College, Aizawl -796005, Mizoram, India
[*]

ABSTRACT – Early detection of Type 2 Diabetes Mellitus using genomic characteristic is inspiring due to large dimensionality, huge missing values, and large number of class imbalance. Bioinformatics has made it easier to analyse complicated biological data, and the fact that an increasing number of this data is available has made machine learning techniques far more beneficial. The study proposes a hybrid system by integrating preprocessing, hybrid oversampling and feature extraction for the Mizoram population. Missing value filtering, K-Nearest Neighbours imputation, and categorical encoding are performed for curating data. The combination of the ADASYN and Edited Nearest Neighbors approach is used to address class imbalance. Feature extraction uses a genetic algorithm to choose useful representations after combining leaf indices from Decision Tree, Random Forest, and XGBoost models. Minority class detection is constantly improved by machine learning models and the hybrid LFGA–ADENN pipeline, reaching up to 0.921 F1-score and 0.920 MCC using XGBoost. LFGA–ADENN shows strong performance in imbalanced genomic T2DM prediction, improving recall while preserving precision when compared to no oversampling or single-method approaches.

Keywords: *T2DM, Genomic Sequencing, Hybrid Feature Extraction, Leaf Fusion, Genetic Algorithm, Hybrid Oversampling.*

1. INTRODUCTION

Type 2 Diabetes Mellitus (T2DM) is becoming the world-wide health issues, covering 589 million individuals around the world and the impact of this disease gradually damaging various human organs such as heart, sight, nerve and kidney (Genitsaridi *et al., 2026*). Effective intervention and management of T2DM depend on early detection. However existing diagnostic techniques frequently rely on clinical assessments that might fail to identify people in pre diabetic or asymptomatic peaks (Tabák *et al., 2012*; American Diabetes Association. 2023).

Since environmental factors, lifestyle changes, and genetic susceptibility, the prevalence of T2DM is increasing in India, especially among the Mizoram people (Brindha *et al., 2022*). Despite the growing availability of genomic sequencing data, predictive modelling for early detection remains underexplored in this population (Martin *et al., 2019*). Most existing studies focus on conventional clinical features or standard machine learning (ML) models, which are often limited by high-dimensional genomic data (Libbrech & William. 2015), severe class imbalance (Chicco & Jurman, 2020), and noisy features, leading to

suboptimal predictive performance (Bolón *et al. 2015*).

To address these challenges, this research proposes hybrid of Random Forest (RF), Leaf Fusion and Genetic Algorithm (LFGA), Adaptive Differential Evolution Neural Network (ADENN). The RF–LFGA–ADENN model is designed to enhance feature extraction and oversampling for early T2DM prediction. The study approach leverages LFGA to capture complex interactions among genomic variants and select the most informative features. Simultaneously, the hybrid ADASYN–ENN (ADENN) approach combines Adaptive Synthetic Sampling (ADASYN) to generate minority class samples with Edited Nearest Neighbours (ENN) to remove noisy instances, effectively balancing the dataset and improving data quality (He *et al.* 2008). The final RF classifier trained on GA optimized fused features enables robust and interpretable early prediction.

This study bridges the research gap in population specific genomic modelling for T2DM, offering a scalable and accurate framework tailored to the Mizoram population.

2. RELATED WORK

T2DM is recognized as one of the most pervasive metabolic disorders worldwide, with rising incidence in both developed and developing countries (Zheng *et al.* 2018). Early prediction of T2DM has become critical to prevent severe complications, reduce healthcare costs, and improve quality of life (Trikkalinou *et al.* 2017). Recent advancements in genomic sequencing and high throughput technologies have enabled large scale profiling of genetic variants associated with T2DM. Genome wide association studies (GWAS) have identified multiple loci linked to disease susceptibility, emphasizing the potential of genomics for early detection (Nasykhova *et al. 2019*). However, translating high dimensional genomic data into accurate predictive models remains a significant computational and statistical challenge due to dimensionality, sparsity, and complex interactions among variants (Wang *et al,* 2022).

While several predictive models exist for T2DM in global populations, region-specific studies remain scarce. Populations in Northeast India, such as Mizoram, exhibit distinct genetic backgrounds and environmental influences that may modulate disease risk. Existing epidemiological studies in Mizoram indicate a rising prevalence of T2DM linked to lifestyle transitions and genetic predisposition. Yet, the application of genomic based predictive modelling in this population is limited, creating a gap in tailored early detection strategies that consider both local genetic diversity and disease dynamics (Lalrohlui *et al.* 2021; Kharsati *et al.* 2024).

High dimensionality is a primary challenge in genomic data analysis (Libbrech & William. 2015). Conventional statistical approaches often fail to capture nonlinear relationships or interactions among genetic variants (Moore *et al.* 2010). ML techniques, such as RF, Decision Trees (DT), and XGBoost (XGB), have shown promise in handling complex genomic datasets, providing both feature importance metrics and predictive power. Nonetheless, feature redundancy, noise, and sparse informative signals can reduce model accuracy (Chen & Hemant. 2016). Hybrid feature extraction methods that combine multiple models or incorporate evolutionary algorithms have emerged to address these challenges, and enabling the identification of highly discriminative genomic patterns (Saeys *et al.* 2007).

Leaf fusion (LF), which aggregates leaf node information from ensemble tree models, has been demonstrated to capture structural patterns and interactions among features that individual trees might overlook (Sirocchi *et al.* 2025). When coupled with a Genetic Algorithm (GA), this approach allows for optimal selection of informative features by

evaluating subsets against predictive performance. The studied conducted using hybrid LFGA approaches boost the model interpretability and accuracy in clinical and genomic data (Xue *et al.* 2015). While its implementation in population specific genome sequencing data for early T2DM prediction is still mostly unexplored, especially in underrepresented populations like Mizoram (Lalrohlui *et al.* 2021).

Class imbalance is a significant barrier to T2DM prediction since, in genomic cohorts, the number of disease cases is frequently lower than that of controls (Krawczyk. 2016). Poor sensitivity may result from standard classifiers bias toward majority classes (He & Garcia, 2009). While ENN and related cleaning approaches minimize noisy or borderline samples, oversampling techniques like SMOTE and ADASYN artificially create minority samples to balance datasets (Chawla *et al.,* 2002). Although hybrid oversampling techniques that combine noise reduction and synthetic generation have been demonstrated to increase model robustness, particularly in high-dimensional genomic contexts, they have not been extensively used in population specific T2DM datasets (Saeys *et al.* 2007).

RF can excellently handle high volume of data like genomic characteristics for classification on account of its resilience to overfitting, robustness, and capacity to manage feature interactions (Breiman. 2001). RF can take advantage of complementary decision patterns across several models when combined with ensemble learning techniques like XGB and DT based LF (Chen & Carlos, 2016). Research shows that ensemble LF frameworks perform better than single model approaches in terms of accuracy and feature interpretability when optimized by GA (Dietterich, 2000, Xue *et al.*, 2015). However, there hasn't been much research done on integrating these techniques with hybrid oversampling in a cohesive pipeline for early T2DM prediction.

Significant gaps still exist despite advancements in T2DM genomic prediction: (i) insufficient population specific research for underrepresented areas such as Mizoram. (ii) poor handling of high dimensional and noisy genomic features. (iii) inadequate integration of hybrid feature extraction and class balancing strategies and (iv) limited use of ensemble based, GA optimized pipelines for early prediction. A system that simultaneously controls class imbalance, obtains discriminative features from several models, and makes use of reliable predictive classifiers is necessary to close these gaps.

In the light of these challenges and in order to provide precise and comprehensible early detection of T2DM in the Mizoram population, the investigation result suggests RF–LFGA–ADENN, which combines hybrid feature extraction (LFGA), hybrid oversampling (ADENN), and RF classification.

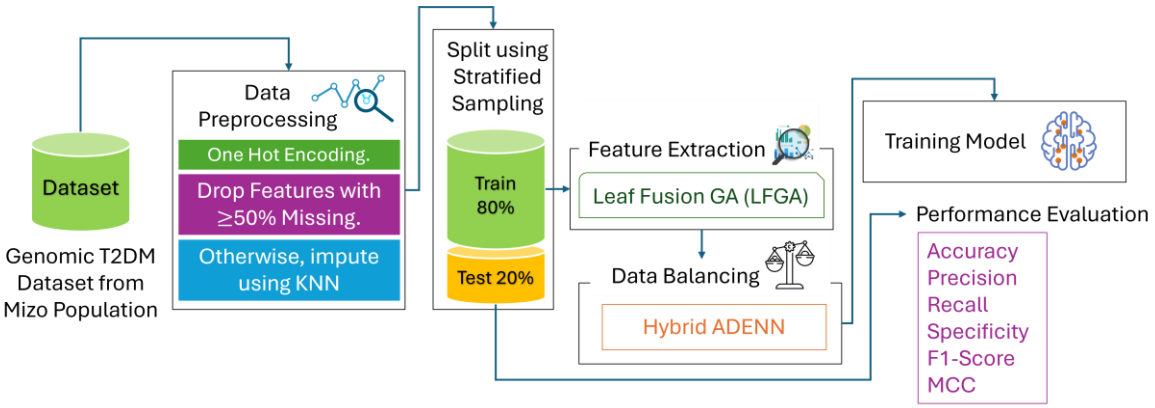

Fig. 1: Proposed Research Flow

3. PROPOSED METHODOLOGY

RF-LFGA-ADENN, a hybrid framework intended for early T2DM prediction from genomic sequencing data in the Mizoram population is proposed in this paper. As shown in Fig. 1, the methodology combines RF classifier model, hybrid over sampling, hybrid feature extraction, and data preparation.

3.1 DATA DESCRIPTION

The study depends on the WES dataset sources from the Biotechnology Department of Mizoram University. The dataset is the combination of 27 non-diabetic and 18 diabetics medically confirmed patients. The diagnosis was verified by a diabetologist in accordance with the World Health Organization's (WHO) 2003 norms. Plasma glucose levels of more than 200 mg/dL after a meal and 120 mg/dL following a fasting period became diagnostic parameters. Most of the people who took part were between 40 and 80 years old, and they all signed a form saying they understood what they were doing. To ensure quality and consistency, all genomic samples were processed and annotated in a standardised way by subject matter experts.

3.2 VARIANT PROCESSING AND WHOLE EXOME SEQUENCE (WES) ANALYSIS:

Genomic DNA was extracted from diabetic patient blood samples (Sarma *et al. 2023*), and WES (100X coverage) was carried out using an Illumina NovaSeq 6000. The data was analysed using an automated tool called WEAP (Ghatak *et al. 2013*). The steps that were taken are displayed below:

Raw FASTQ reads were subjected to quality control (FastQC) (Yang *et al.* 2013) and trimming (Trimmomatic) (Bolger *et al. 2014*) prior to alignment to the reference genome using BWA or Stampy and the generation of SAM files (Lunter & Martin. 2011). File Conversion & Processing: SAM files were converted to sorted BAM format using Samtools (Li *et al.,* 2009), and duplicate reads were found using Picard (Picard Toolkit. 2019). Variant Calling: VCF files with indels and SNPs were exported (Rimmer *et al.* 2014). Annotation: ClinVar and dbSNP databases were consulted in order to use Annovar to annotate variants for clinical and functional significance (Wang *et al.* 2010). Filtering: Pathogenic variants were ranked according to gnomAD population frequencies and functional scores, while low-quality variants were eliminated (Ortiz *et al. 2024*).

The annotated VCF files included minor allele frequencies across all geographic populations, along with functional impact predictions for each single nucleotide polymorphism (SNP). These predictions were obtained from multiple computational tools, including Combined Annotation Dependent Depletion (CADD) (Rentzsch *et al.,* 2019), Polymorphism Phenotyping v2 (Polyphen2) (Adzhubei *et al.,* 2013), Functional Analysis through Hidden Markov Models (FATHMM) (Shihab *at el., 2013*). Variant Effect Scoring Tool V3 (VEST3) (Douville *et al.,* 2016), Sorting Intolerant from Tolerant (SIFT) (Sim *et al., 2012*), Likelihood Ratio Test (LRT) (Zeng *et al, 2014*), MutationAssessor (Reva *et al.,* 2011), MutationTaster and RadialSVM (Castellana *et al., 2013*). Additionally, several manually curated features were incorporated, such as heterogeneous and homogeneous SNVs, amino acid alteration SNVs, exon splicing SNVs, start-loss and stop-loss SNVs, as well as chromosome number, gene name, reference and alternate bases, and positional information.

3.4 DATA PREPROCESSING

Data preprocessing is a key component of the medical data classification process, information analysis, and ML pipeline. Medical data often demonstrates missing values (García *et al.* 2010), imbalanced classes (He *et al*. 2009), high-dimensional features (Saeys *et al*. 2007), and the removal of duplicate or redundant data (Cios *et al*.

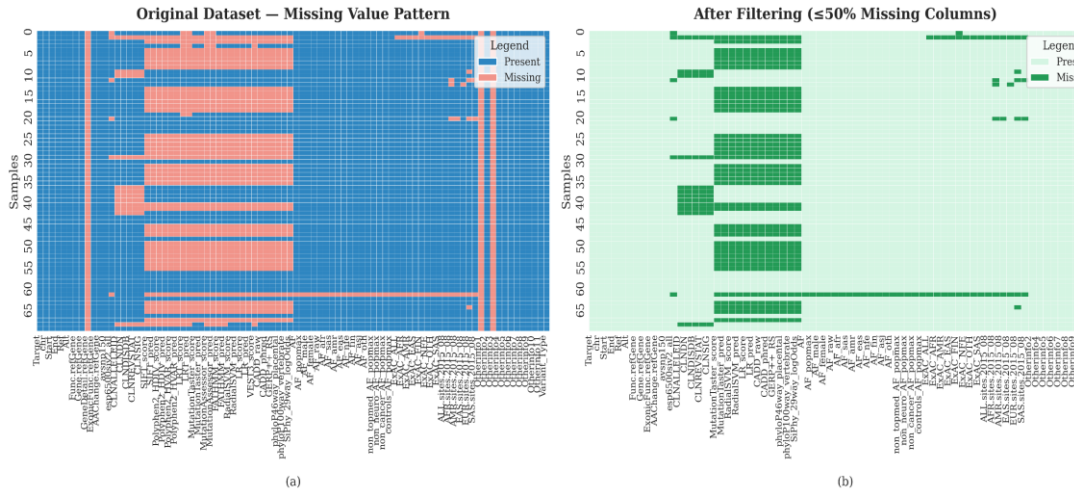

Fig. 2: Missing value patterns: (a) before, and (b) after dropping features with ≥ 50% missing values.

2002); thus, appropriate preprocessing can substantially enhance model performance. The first and most critical step in the preprocessing pipeline is ensuring data quality by removing features with missing values exceeding a predefined threshold (50%), as illustrated in Fig. 2, showing the dataset before (a) and after (b) filtering missing features. Retaining highly incomplete features may introduce bias and adversely affect model reliability.

For the remaining missing values, we have tested several imputation strategies, including mean, median, KNN, and iterative imputation, to identify the optimal method. As shown in Fig. 3, KNN imputation yielded the best predictive performance on a 20% test split and was therefore employed for all downstream analyses. The data is processed using the algorithm 1 to have curated data.

Algorithm 1 Data Preprocessing for Genomic T2DM Dataset
Input: Genomic dataset *D* with features *X* and target *y*
Output: Preprocessed training and test sets *(Xtrain, Xtest, ytrain, ytest)*
1. **For** each feature $f \in X$, **do:**
2. **If** missing percentage *(f)* ≥ *50%*, **then:**
3. Drop feature *f*
4. **Else:**
5. Impute missing values using KNN
6. **End If**
7. **End for**
8. Identify categorical columns *Ccat* and numerical columns *Cnum*
9. Encode *Ccat* using One-Hot Encoding → *Xcat_ohe* 10. Combine numerical and encoded features: *Xprocessed = [Xnum, Xcat_ohe]*
11. Split $Xprocessed$ and y into 80% train and 20% test: *(Xtrain, Xtest, ytrain, ytest)* ← *TrainTestSplit(Xprocessed, y, 0.8)*

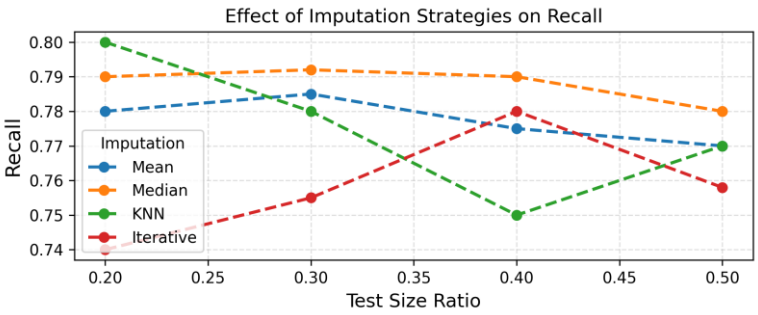

Fig. 3: Compared Recall for different test set ratios using different imputation techniques.

In order to ensure compatibility with tree based models and prevent unintentional ordinal associations, categorical characteristics were then converted to numeric form using one hot encoding. Numerical features were then standardized as needed to allow distance based computations during hybrid oversampling. The LFGA feature extraction was then used prior to any resampling. By verifying that oversampling is performed in a condensed and informative feature space, this step lowers the possibility of producing redundant or noisy synthetic samples. Hybrid oversampling (ADENN), which successively integrates ADASYN and ENN, was then used to address class imbalance. ADASYN adaptively creates synthetic minority samples in locations that are difficult to learn, while ENN eliminates noisy or incorrectly categorized cases close to class boundaries. As seen in Fig. 4, this ordering LFGA followed by ADENN guarantees that oversampling is directed by optimized features and that the final training set is both balanced and noise-reduced.

3.4.1 HYBRID LEAF FUSION GENETIC ALGORITHM (LFGA)

The study uses hybrid feature extraction utilizing LFGA to efficiently manage high dimensional genomic data and capture intricate interactions among variations. First, a GA is applied on the raw features to remove irrelevant or redundant dimensions, selecting

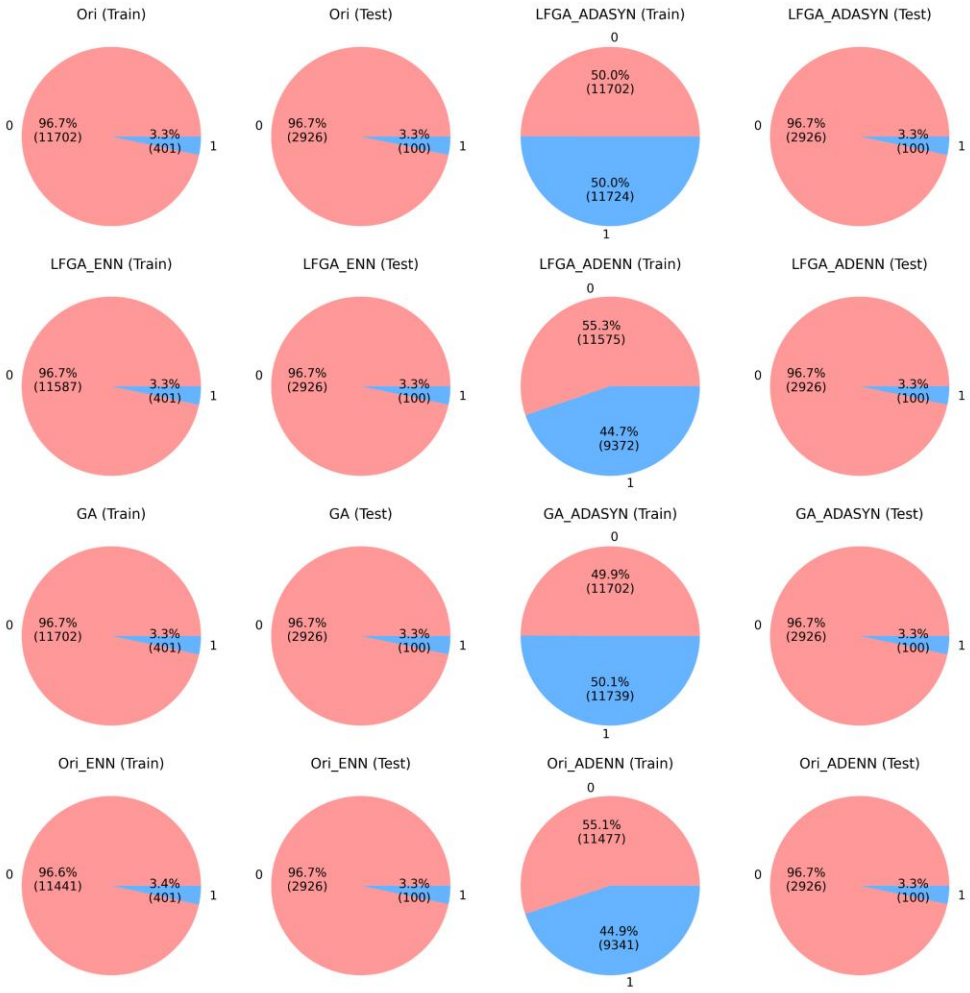

Fig. 4: Class distribution of the original (Ori) dataset (train and test) and resampling strategies on the training set using GA, and LFGA feature sets.

a subset of highly informative features for subsequent modelling. Next, DT, RF, and XGB models are trained on the selected raw features to exploit complementary decision patterns.

The leaf indices from all three models are then fused into a combined feature matrix, encoding how samples traverse different decision paths and preserving hierarchical and interaction information that raw features alone may miss. This fused matrix is one hot encoded to make categorical leaf information usable for downstream modelling. GA is subsequently used on the fused features to select the most discriminative leaf-based patterns, retaining only the most informative fused features. The LFGA feature extraction flow is illustrated in Fig. 5, while the complete details are provided in Algorithm 2.

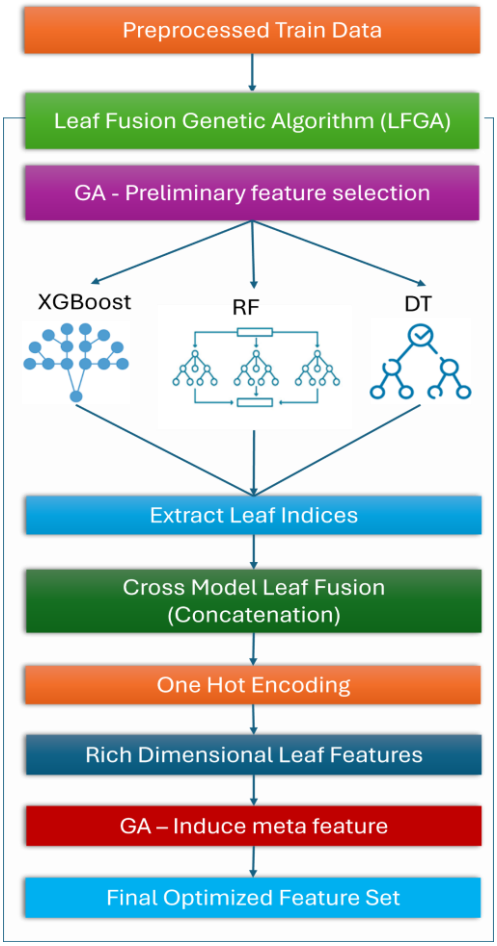

Fig. 5: Leaf Fusion-Based LFGA feature Extraction

Algorithm 2 LFGA Feature Extraction	
1	**Input:** Preprocessed training set X_{train}, Y_{train}) and test set X_{test}
2	**Output:** Optimized raw features *(Xtrain raw sel, Xtest raw sel)* and fused features *(Xtrain fused sel, Xtest fused sel)*
3	**Step 1: GA on Raw Features**
4	Initialize population of binary masks: **M = {m₁, m₂, ..., m_{npop}}**
5	for each generation **g = 1 to n_gen** do
6	for each mask **$m_i \in M$** do
7	Select features: **$X^{train(i)}$ = Xtrain[:, m_i = 1]**
8	Evaluate fitness using cross-validated F1-score with RF
9	end for
10	Update **M** using selection, crossover, and mutation
11	end for
12	Select best mask **m∗: Xtrain raw sel = Xtrain[:, m∗ = 1], Xtest raw sel = Xtest[:, m∗ = 1]**
13	**Step 2: Leaf Fusion from Ensemble Models**
14	Train models: **DT, RF, XGB**
15	Extract leaf indices: $L^{DT}\_train = DT.apply(Xtrain)$ $L^{Rf}\_train = RF.apply(Xtrain)$ $L^{xGB}\_train = XGB.apply(Xtrain)$
16	Fuse leaves: $L^{fused}\_train = [L^{DT}\_train, L^{Rf}\_train, L^{xGB}\_train]$
17	One-hot encoding: $Xtrain\ fused, Xtest\ fused$
18	**Step 3: GA on Fused Feature Space**
19	Apply GA: $Xtrain\ fused\ sel = Xtrain\ fused[:, m*^{f} = 1], Xtest\ fused\ sel = Xtest\ fused[:, m*^{f} = 1]$
20	**Return:** $(Xtrain\ raw\ sel, Xtest\ raw\ sel)$ and $(Xtrain\ fused\ sel, Xtest\ fused\ sel)$

3.5 HYBRID OVERSAMPLING (ADENN)

With far fewer positive (confirm diabetes) cases than negative (non-diabetes) samples, genomic datasets for T2DM are frequently severely unbalanced. The stability of tree-based classifiers, sensitivity for early detection, and model bias toward the majority class can all be impacted by such an imbalance (He & Garcia, 2009; Chawla *et al.*, 2002). In order to overcome these difficulties, we suggest a hybrid oversampling technique called ADENN, which takes advantage of the complimentary advantages of both adaptive synthetic sampling (ADASYN) and ENN.

3.5.1 ADASYN

By concentrating on hard-to-learn areas of the feature space, where minority cases are surrounded by majority neighbors, ADASYN creates artificial minority samples (He *et al.,* 2008). The Eqn. (1) defines the the difficulty ratio for each minority instance x_i :

$$\Delta_i = \frac{|\{x_{ij} \in k\text{-NN}(x_i): y_{ij} \neq y_i\}|}{k}, \quad i = 1, \dots, N_{min} \qquad (1)$$

The normalized importance of x_i in producing synthetic samples is shown in Eqn. (2):

$$r_i = \frac{\Delta_i}{\sum_{j=1}^{N_{min}} \Delta_j}, \quad i = 1, \dots, N_{min} \qquad (2)$$

Thus, the values of x_i is calculated using Eqn. (3):

$$G_i = r_i \cdot G \qquad (3)$$

where G is the total quantity of synthetic samples needed. ADASYN improves minority representation and lessens classifier bias by adaptively producing extra samples for challenging cases.

3.5.2 ENN

Even if ADASYN balances the dataset, noise may be introduced by certain outputs or original samples which are close to class boundaries (Fernández *et al.,* 2018). Using Eqn. (4), ENN eliminates each instance x_i whose class label varies from the majority label among its k nearest neighbors,

$$x_i \in D \text{ is removed if } y_i \neq \text{mode}\{y_j : x_j \in k\text{-NN}(x_i)\}, \quad i = 1, \dots, N \qquad (4)$$

In genomic data, when minority patterns are rare and noisy, ENN increases border clarity by eliminating noisy or incorrectly categorized samples (Wilson, 1972).

3.5.3 HYBRID ADENN

The strengths of ADASYN and ENN are combined in a sequential but supportive way in the proposed hybrid technique, ADENN. Based on the observed minority ratio, the ADASYN component adaptively creates synthetic minority samples, concentrating on instances that are more difficult to learn (i.e., surrounded by majority neighbors) (He *et al.,* 2008). Then, every instance whose label differs from the majority of its k nearest neighbors—including possibly noisy synthetic points—is eliminated by the ENN component (Wilson, 1972). Balanced minority representation, noise reduction, preservation of informative patterns, compatibility with high-dimensional data, and mitigation of overfitting are all guaranteed by this sequential combination (Krawczyk, 2016).

The Eqn. (5) defines the hybrid process:

$$(X^*, y^*) = ENN\big(ADASYN(X, y)\big) \qquad (5)$$

where (X^*, y^*) is the final training set that is informative, balanced, and minimizing noise at the same time.

The ADENN method on an unbalanced dataset with an initial 9:1 class imbalance (450 majority vs. 50 minority samples) is shown in Fig. 6. The final dataset following ADENN is displayed in the main panel (a), with majority points (blue) and minority points classified as easy-to-learn (light orange) and hard-to-learn (dark orange).

In this case, hard-to-learn points are at decision boundaries and close to majority samples, making them more difficult to learn, whereas easy-to-learn points are minority occurrences far from majority samples, indicating areas where the classifier can consistently learn. Inset (b) illustrates how ADASYN concentrates on challenging areas by highlighting ADASYN-generated synthetic points (green) with gray dashed lines connecting them to the minority neighbors used for generation. ENN noise filtering is shown in inset (c), where the original minority and majority positions prior to ENN are indicated by red crosses. This illustration

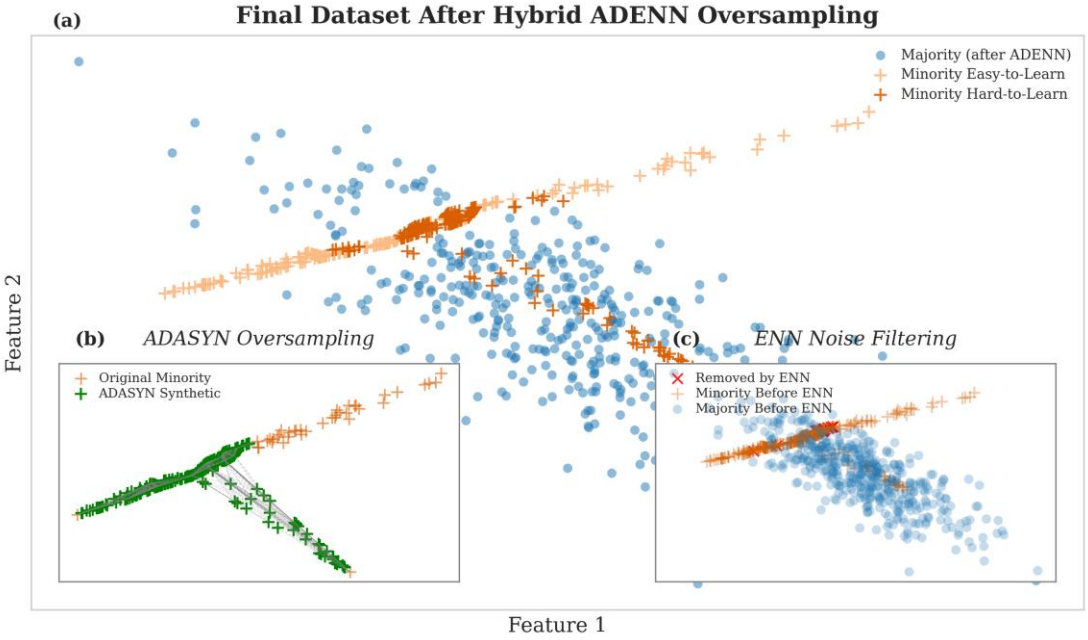

Fig. 6. A step-by-step demonstration of the Hybrid LFGA–ADENN procedure using an imbalanced toy dataset. (a) Main panel: final dataset using ADENN, with minority points categorized as easy-to-learn (light orange) and hard-to-learn (dark orange) and majority points (blue). (b) Inset: Oversampling of challenging minority regions is highlighted by ADASYN synthetic points (green) with grey dashed lines indicating neighbors used for generation. (c) Inset: ENN noise filtering, displaying minority (orange) and majority (blue) points before to ENN along with points eliminated

shows how the ADENN pipeline eliminates noisy samples, maintains minority structure, and balances the dataset.

3.6. MODEL TRAINING

For early detection of T2DM using genomic characteristics, this study trained couples of tree-based classifiers on different feature sets and oversampling techniques.

3.6.1 DECISION TREE (DT):

DT recursively divide the feature space into homogeneous subsets by reducing impurity measures like the Gini index or entropy (Breiman *et al.,* 1984). They are easy to understand and use as baseline classifiers because they have clear rules for making decisions (Quinlan, 1993). Nonetheless, decision trees are susceptible to overfitting, especially when utilised with high dimensional genomic data (Hastie *et al.,* 2009). Even with this flaw, they are still useful for finding important features and are basic parts of more advanced ensemble methods (Breiman, 2001).

3.6.2 RANDOM FOREST (RF):

RF is a bagging-based ensemble of multiple DT trained on boot strapped subsets of the data. Each tree votes, and the majority vote determines the final class, improving stability and reducing variance. RF is robust to high dimensional, sparse, and noisy features, which are common in genomic datasets (Breiman, 2001). It operates as the foundation of our final model, RF–LFGA–ADENN, and is appropriate for early disease prediction due to its capacity to capture hierarchical interactions while maintaining interpretability.

3.6.3 XGBoost (XGB):

In order to avoid overfitting, XGB is a gradient boosting framework that generates trees one after the other while optimizing residual errors with regularization at each

stage. It effectively manages high dimensional, sparse data, capturing intricate feature interactions. For speed and accuracy, XGB employs advanced approaches like parallel processing, learning rate adjustment, and tree pruning (Chen & Guestrin, 2016). In biomedical data, where subtle feature interactions are crucial, this model works very well. Every model was assessed using several oversampling techniques, such as no oversampling, ADASYN, ENN, and hybrid ADENN, after being trained on several feature sets, including original, GA-selected, and leaf-fusion features. The greatest results were obtained by RF trained on GA and LF features with hybrid oversampling (ADENN), which is referred to as RF–LFGA–ADENN.

3.7. PERFORMANCE EVALUATION METRICS

Several commonly used classification metrics were employed to assess the performance of the suggested models, including RF-LFGA-ADENN. These metrics offer a thorough evaluation, especially for genomic T2DM datasets that are unbalanced. The evaluation metrics, which are computed using Eqn. (6) to (11), including Accuracy, Precision, F1-score, Matthew's correlation coefficient (MCC), Recall, and Area under the precision-recall curve (AUC-PR). These metrics provide a comprehensive understanding of discriminative power and class balance sensitivity (Kahn *et al.,* 2000).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \quad (9)$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (10)$$

$$AUC\text{-}PR = \int Precision \; d(Recall) \quad (11)$$

4. RESULT AND DISCUSSION

The experimental findings obtained using baseline classifiers under different feature extraction and oversampling strategies. Table 1 reports base line performance on the original imbalanced dataset without resampling.

Table 2:Performance Comparison of Baseline Models on Original Data

Model	Accuracy	Precision	Recall	Specificity	F1 Score	MCC
DT	**0.994**	**0.994**	**0.994**	0.998	**0.994**	**0.910**
RF	0.991	0.990	0.991	**0.998**	0.990	0.846
XGB	**0.994**	**0.994**	**0.994**	0.998	**0.994**	**0.910**

Table 1: Performance of Baseline DT Across Different Sampling Techniques

Dataset	Accuracy	Precision	Recall	Specificity	F1-Score	MCC
LFGA–ADENN	0.993	0.857	**0.960**	0.995	0.906	0.904
GA ADENN	0.992	0.858	0.910	0.995	0.883	0.880
GA ENN	0.988	0.764	0.940	0.990	0.843	0.842
GA	0.995	0.939	0.920	0.998	**0.929**	**0.927**
LFGA ENN	0.992	0.880	0.880	0.996	0.880	0.876
LFGA ADASYN	**0.995**	**0.978**	0.880	**0.999**	0.926	0.925
GA ADASYN	0.993	0.916	0.870	0.997	0.892	0.889

Although high accuracy values are observed MCC values remain relatively moderate (0.910), indicating that accuracy alone is insufficient for evaluating performance under severe imbalance.

This observation motivates the use of oversampling and feature extraction

strategies. The impact of the proposed LFGA feature extraction is evident in Table 2, Table 3, and Table 4.

For the DT classifier Table 2, LFGA–ADENN improves recall to 0.960 and F1-score to 0.906, compared to GA-only configurations

Table 4: Performance of Baseline XGB Across Different Sampling Techniques

Dataset	Accuracy	Precision	Recall	Specificity	F1-Score	MCC
LFGA–ADENN	**0.994**	0.861	**0.990**	0.995	0.921	0.920
GA ADENN	0.993	0.857	0.960	0.995	0.906	0.904
GA ENN	0.987	0.746	0.940	0.989	0.832	0.831
GA	**0.994**	0.918	0.890	0.997	0.904	0.900
LFGA ENN	0.991	0.800	0.960	0.992	0.873	0.872
LFGA ADASYN	0.994	**0.937**	0.890	**0.998**	0.913	0.910
GA ADASYN	**0.994**	0.919	0.910	0.997	**0.915**	**0.912**

Table 3: Performance of Baseline RF Across Different Sampling Techniques

Dataset	Accuracy	Precision	Recall	Specificity	F1-Score	MCC
LFGA–ADENN	0.993	0.836	**0.970**	0.994	0.898	0.897
GA ADENN	0.990	0.832	0.890	0.994	0.860	0.855
GA ENN	0.982	0.685	0.850	0.987	0.759	0.754
GA	0.991	0.929	0.790	**0.998**	0.854	0.852
LFGA ENN	0.992	0.860	0.920	0.995	0.889	0.886
LFGA ADASYN	**0.994**	**0.928**	0.900	**0.998**	**0.914**	**0.911**
GA ADASYN	0.992	0.913	0.840	0.997	0.875	0.872

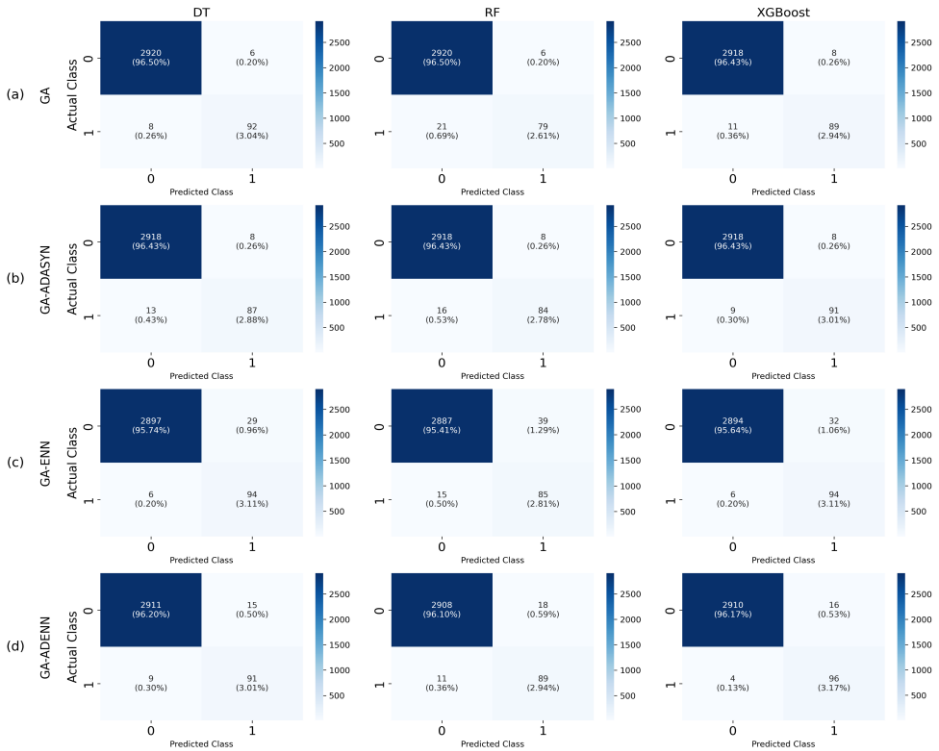

Fig. 7: Confusion matrices of three classifiers evaluated on GA selected features with different oversampling strategies: (a) GA based feature selection only, (b) GA features with ADASYN oversampling, (c) GA features with ENN oversampling, and (d) GA features with hybrid ADENN oversampling.

(F1-score 0.929) and conventional method oversampling such as GA–ENN (F1-score 0.843). Similarly, for RF (Table 3), LFGA–ADENN achieves a recall of 0.970 and F1-score of 0.898, outperforming GA–ADENN (F1-score 0.860) and GA–ENN (F1-score 0.759). These results demonstrate that combining adaptive oversampling with noise filtering improves minority class recognition. XGB results in Table 2 further support this trend. LFGA–ADENN yields the highest recall of 0.990 and F1-score of 0.921, compared to GA–ADENN (F1-score 0.906) and GA–ENN (F1-score 0.832).

more distinct class separation across all baseline classifiers when compared to GA-only and conventional oversampling techniques. This demonstrates that hybrid feature extraction and hybrid oversampling, rather than classifier-specific bias, are the main sources of performance gains.

Overall, the discussion shows that LFGA–ADENN enhances the balance between precision, recall, and MCC across baseline classifiers; DT and RF consistently outperform their baseline configurations, while XGB achieves the greatest absolute scores.

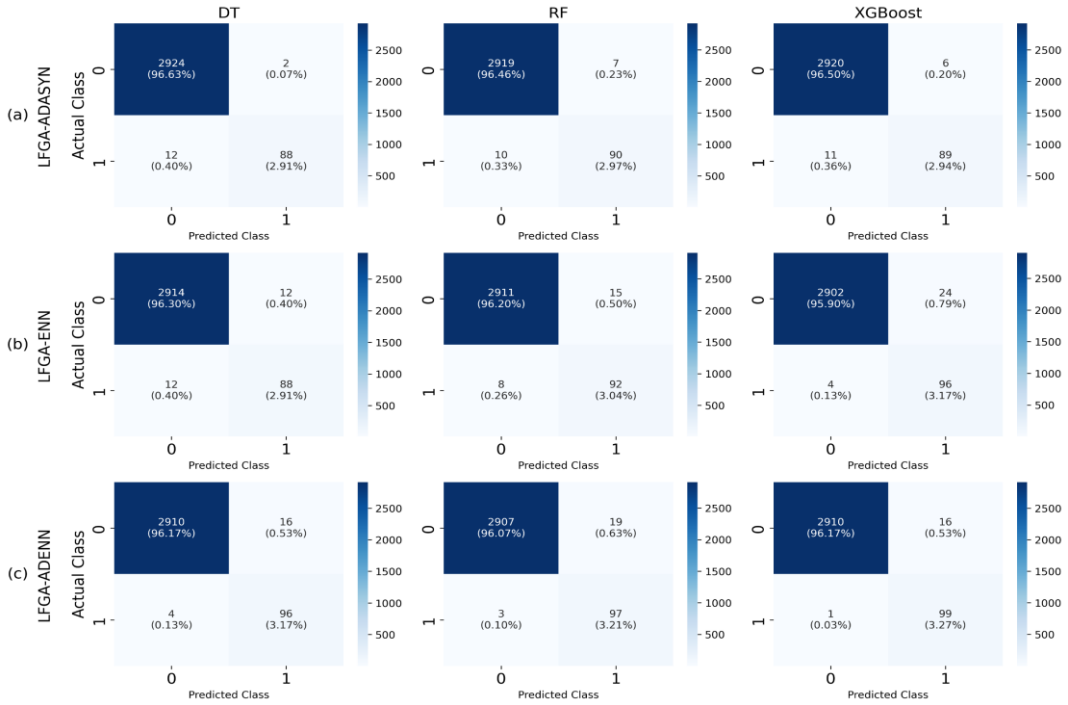

Fig. 8: Confusion matrices of LFGA selected features with different oversampling strategies: (a) LFGA features with ADASYN oversampling, (b) LFGA features with ENN oversampling, and (c) LFGA features with hybrid ADENN oversampling.

Improved balanced classification is confirmed by the related MCC of 0.920. With great specificity (≥ 0.994) and lower false negatives without excessive false positives, LFGA–ADENN consistently delivers higher recall across all classifiers.

These numerical results are graphically supported by the confusion matrices shown in Fig. 7 and 8. The LFGA–ADENN configuration shows fewer false negatives and

CONCLUSION

This work used imbalanced genomic sequencing data from the Mizoram population to assess a hybrid learning pipeline that combines ADENN-based hybrid oversampling and LFGA based feature extraction for early T2DM prediction. The suggested method was evaluated under consistent experimental conditions employing baseline classifiers, such as DT, RF, and XGB.

The experimental outcome shows that minority class detection is consistently improved by the LFGA–ADENN combination across all classifiers. The highest performance is recorded by XGB with an F1-score of 0.921 and MCC of 0.920, followed by Decision Tree with an F1-score of 0.906 and MCC of 0.904, RF with an F1-score of 0.898 and MCC of 0.897. Remarkably, recall values rise to 0.96–0.99 with LFGA–ADENN, suggesting a significant decrease in false negatives when compared to single-method oversampling and no oversampling techniques.

These results verify that improves at data as well as feature thresholds, rather than depending on a particular classifier, are what cause performance gains. For imbalanced genomic prediction tasks, the combination of adaptive oversampling, noise filtering, and leaf fusion-based feature representation offers a reliable and broadly applicable solution. Future research will concentrate on expanding the framework to include clinical characteristics in addition to genomic traits and verifying it on bigger, independent cohorts.

AUTHORS CONTRIBUTION

Vanlalawmpuia R and Lalhmingliana were the primary researchers responsible for formulating articles and processing data. Brindha Senthil Kumar, Freda Lalrohlui, Vanlalhruaii, John Zohmingthanga contributed to data collection and background development. Nachmimuthu Senthil Kumar and Lalhmingliana played a key role in designing the learning framework and facilitating discussions.

FUNDING

Not Applicable

ACKNOWLEDGEMENT

The Department of Biotechnology, New Delhi, provided infrastructure support through Mizoram University's Advanced Level State Biotech Hub (BT/NER/143/SP44475/2021), which was crucial to the successful completion of this study. For this, the authors are truly grateful. The researchers also thank the lab technicians and academics from Mizoram University's Departments of Biotechnology and Computer Engineering for their invaluable support and collaboration during the study

CONFLICT OF INTEREST

The authors declare there is no conflict of interest.

ETHICAL STATEMENT

The Human Ethical Committee of Mizoram University (MZU/IHEC/2015/006 dtd. 14/12/15) and the Civil Hospital, Aizawl (B.12018/1/13-CH (A)IEC/39 dtd. 23/12/2015) have both approved the study protocol.

REFERENCES

Adzhubei, Ivan, Daniel M. Jordan, and Shamil R. Sunyaev. "Predicting functional effect of human missense mutations using PolyPhen-2." *Current protocols in human genetics* 76, no. 1 (2013): 7-20. https://doi.org/10.1002/0471142905.hg0720s76

Alonso-Betanzos, Amparo, and Verónica Bolón-Canedo. "Big-data analysis, cluster analysis, and machine-learning approaches." *Sex-specific analysis of cardiovascular function* (2018): 607-626. *https://doi.org/10.1007/978-3-319-77932-4_37*

American Diabetes Association. "Standards of care in diabetes—2023 abridged for primary care providers." *Clinical Diabetes* 41, no. 1 (2023): 4-31. https://doi.org/10.2337/cd23-as01

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30, no. 15 (2014): 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Bolón-Canedo, Verónica, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. "Recent advances and emerging challenges of feature selection in the context of big data." *Knowledge-based systems* 86 (2015): 33-45. https://doi.org/10.1016/j.knosys.2015.05.014

Breiman, Leo, Jerome Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees*. Chapman and Hall/CRC, 2017. https://doi.org/10.1201/9781315139470

Breiman, Leo. "Random forests." *Machine learning* 45, no. 1 (2001): 5-32. https://doi.org/10.1023/A:1010933404324

Brindha Senthil Kumar, Vanlalawmpuia R, Freda Lalrohlui, John Zohmingthanga, Lalruatpuii Hlawnmual, Nachimuthu Senthil Kumar and Lal Hmingliana. "A Multilayer Perceptron Model to Predict Risk Factors of Type II Diabetes Mellitus". *Int J Food Nutr*, 11 (2022): 67-74. https://DOI:10.4103/ijfans_110-22

Castellana, Stefano, and Tommaso Mazza. "Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools." *Briefings in bioinformatics* 14, no. 4 (2013): 448-459. https://doi.org/10.1093/bib/bbt013

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357. https://doi.org/10.1613/jair.953

Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794. 2016. https://doi.org/10.1145/2939672.2939785

Chen, Xi, and Hemant Ishwaran. "Random forests for genomic data analysis." *Genomics* 99, no. 6 (2012): 323-329. https://doi.org/10.1016/j.ygeno.2012.04.003

Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." *BMC genomics* 21, no. 1 (2020): 6. https://doi.org/10.1186/s12864-019-6413-7

Cios, Krzysztof J., Witold Pedrycz, and Roman W. Swiniarski. Data mining methods for knowledge discovery. Springer Science & Business Media, 2012.

Dietterich, Thomas G. "Ensemble methods in machine learning." In *International workshop on multiple classifier systems*, pp. 1-15. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000. https://doi.org/10.1007/3-540-45014-9_1

Douville, Christopher, David L. Masica, Peter D. Stenson, David N. Cooper, Derek M. Gygax, Rick Kim, Michael Ryan, and Rachel Karchin. "Assessing the pathogenicity of insertion and deletion variants with the variant effect scoring tool (VEST-Indel)." *Human mutation* 37, no. 1 (2016): 28-35. https://doi.org/10.1002/humu.22911

Fernández, Alberto, Salvador Garcia, Francisco Herrera, and Nitesh V. Chawla. "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary." *Journal of artificial intelligence research* 61 (2018): 863-905. https://doi.org/10.1613/jair.1.11192

Genitsaridi, Irini, Paraskevi Salpea, Agus Salim, Seyedeh Forough Sajjadi, Dunya Tomic, Steven James, Sathish Thirunavukkarasu et al. "of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050." *The Lancet Diabetes & Endocrinology* 14, no. 2 (2026): 149-156. https://doi.org/10.1016/S2213-8587(25)00299-2

Ghatak, Souvik, Rajendra Bose Muthukumaran, and Senthil Kumar Nachimuthu. "A simple method of genomic DNA extraction from human samples for PCR-RFLP analysis." *Journal of biomolecular techniques: JBT* 24, no. 4 (2013): 224. https://doi.org/10.7171/jbt.13-2404-001

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. "The Elements of Statistical Learning, (2nd printing ed.)." (2009).

He, Haibo, and Edwardo A. Garcia. "Learning from imbalanced data." *IEEE Transactions on knowledge and data engineering* 21, no. 9 (2009): 1263-1284. https://doi.org/10.1109/TKDE.2008.239

He, Haibo, Yang Bai, Edwardo A. Garcia, and Shutao Li. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1322-1328. Ieee, 2008. https://doi.org/10.1109/IJCNN.2008.4633969

Kahn, Barbara B., and Jeffrey S. Flier. "Obesity and insulin resistance." The Journal of clinical investigation 106, no. 4 (2000): 473-481. https://doi.org/10.1172/JCI10842

Kharsati, Naphisabet, and Mrinmoyi Kulkarni. "Living with diabetes in Northeast India: An exploration of psychosocial factors in management." *Dialogues in Health* 4 (2024): 100180. https://doi.org/10.1016/j.dialog.2024.100180

Krawczyk, Bartosz. "Learning from imbalanced data: open challenges and future directions." *Progress in artificial intelligence* 5, no. 4 (2016): 221-232. https://doi.org/10.1007/s13748-016-0094-0

Lalrohlui, Freda, John Zohmingthanga, Andrew Vanlallawma, and Nachimuthu Senthil Kumar. "Whole exome sequencing identifies the novel putative gene variants related with type 2 diabetes in Mizo population, northeast India." *Gene* 769 (2021): 145229. https://doi.org/10.1016/j.gene.2020.145229

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. "The sequence

alignment/map format and SAMtools." *bioinformatics* 25, no. 16 (2009): 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Libbrecht, Maxwell W., and William Stafford Noble. "Machine learning applications in genetics and genomics." *Nature Reviews Genetics* 16, no. 6 (2015): 321-332. https://doi.org/10.1038/nrg3920

Lunter, Gerton, and Martin Goodson. "Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads." *Genome research* 21, no. 6 (2011): 936-939. http://www.genome.org/cgi/doi/10.1101/gr.111120.110

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. "Clinical use of current polygenic risk scores may exacerbate health disparities." *Nature genetics* 51, no. 4 (2019): 584-591. https://doi.org/10.1038/s41588-019-0379-x

Moore, Jason H., Folkert W. Asselbergs, and Scott M. Williams. "Bioinformatics challenges for genome-wide association studies." *Bioinformatics* 26, no. 4 (2010): 445-455. https://doi.org/10.1093/bioinformatics/btp713

Nasykhova, Yulia A., Yury A. Barbitoff, Elena A. Serebryakova, Dmitry S. Katserov, and Andrey S. Glotov. "Recent advances and perspectives in next generation sequencing application to the genetic research of type 2 diabetes." *World journal of diabetes* 10, no. 7 (2019): 376. https://doi.org/10.4239/wjd.v10.i7.376

Ortiz, Bengie L., Vibhuti Gupta, Rajnish Kumar, Aditya Jalin, Xiao Cao, Charles Ziegenbein, Ashutosh Singhal, Muneesh Tewari, and Sung Won Choi. "Data preprocessing techniques for AI and machine learning readiness: Scoping review of wearable sensor data in cancer care." *JMIR mHealth and uHealth* 12, no. 1 (2024): e59587. https://doi.org/10.2196/59587

Toolkit, Picard. "GitHub repository." *Broad Institute. Available online at: http://broadinstitute.github.io/picard* (2019)

Quinlan, J. Ross. *C4. 5: programs for machine learning*. Elsevier, 2014.

Rentzsch, Philipp, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher. "CADD: predicting the deleteriousness of variants throughout the human genome." *Nucleic acids research* 47, no. D1 (2019): D886-D894. https://doi.org/10.1093/nar/gky1016

Reva, Boris, Yevgeniy Antipin, and Chris Sander. "Predicting the functional impact of protein mutations: application to cancer genomics." *Nucleic acids research* 39, no. 17 (2011): e118-e118. https://doi.org/10.1093/nar/gkr407

Rimmer, Andy, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen RF Twigg, WGS500 Consortium, Andrew OM Wilkie, Gil McVean, and Gerton Lunter. "Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications." *Nature genetics* 46, no. 8 (2014): 912-918. https://doi.org/10.1038/ng.3036

Saeys, Yvan, Inaki Inza, and Pedro Larranaga. "A review of feature selection techniques in bioinformatics." *bioinformatics* 23, no. 19 (2007): 2507-2517. https://doi.org/10.1093/bioinformatics/btm344

Sarma, Ranjan Jyoti, Jeremy Lalrinsanga Pautu, Bawitlung Zothankima, Lalfakzuala Khenglawt, Saia Chenkual, John Zohmingthanga, Lalawmpuii Pachuau, and Nachimuthu Senthil Kumar. "Novel germline variants of MUC3A in a patient with ER+ breast cancer and signet-ring cell stomach adenocarcinoma." *Gene Reports* 33 (2023): 101803. https://doi.org/10.1016/j.genrep.2023.101803

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., Day, I. N., & Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Human Mutation, 34(1), 57–65. https://doi.org/10.1002/humu.22225

Sim, Ngak-Leng, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C. Ng. "SIFT web server: predicting effects of amino acid substitutions on and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *biorxiv*, 531210. https://doi.org/10.1093/nar/gks539

Sirocchi, Christel, Martin Urschler, and Bastian Pfeifer. "Feature graphs for interpretable unsupervised tree ensembles: centrality, interaction, and application in disease subtyping." *BioData Mining* 18, no. 1 (2025): 15. https://doi.org/10.1186/s13040-025-00430-3

Tabák, Adam G., Christian Herder, Wolfgang Rathmann, Eric J. Brunner, and Mika Kivimäki. "Prediabetes: a high-risk state for diabetes development." *The Lancet* 379, no. 9833 (2012): 2279-2290. https://doi.org/10.1016/S0140-6736(12)60283-9

Trikkalinou, Aikaterini, Athanasia K. Papazafiropoulou, and Andreas Melidonis. "Type 2 diabetes and quality of life." *World journal of diabetes* 8, no. 4 (2017): 120. https://doi.org/10.4239/wjd.v8.i4.120

Wang, Kai, Mingyao Li, and Hakon Hakonarson. "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." *Nucleic acids research* 38, no. 16 (2010): e164-e164. https://doi.org/10.1093/nar/gkq603

Wang, Tao, Yongzhuang Liu, Quanwei Yin, Jiaquan Geng, Jin Chen, Xipeng Yin, Yongtian Wang et al. "Enhancing discoveries of molecular QTL studies with small sample size using summary statistic imputation." *Briefings in bioinformatics* 23, no. 1 (2022): bbab370. https://doi.org/10.1093/bib/bbac139

Wilson, Dennis L. "Asymptotic properties of nearest neighbor rules using edited data." *IEEE Transactions on Systems, Man, and Cybernetics* 3 (2007): 408-421. https://doi.org/10.1109/TSMC.1972.4309137

Xue, Bing, Mengjie Zhang, Will N. Browne, and Xin Yao. "A survey on evolutionary computation approaches to feature selection." *IEEE Transactions on evolutionary computation* 20, no. 4 (2015): 606-626. https://doi.org/10.1109/TEVC.2015.2504420

Yang, Xi, Di Liu, Fei Liu, Jun Wu, Jing Zou, Xue Xiao, Fangqing Zhao, and Baoli Zhu. "HTQC: a fast quality control toolkit for Illumina sequencing data." *BMC bioinformatics* 14, no. 1 (2013): 33. https://doi.org/10.1186/1471-2105-14-33

Zeng, Ping, Yang Zhao, Jin Liu, Liya Liu, Liwei Zhang, Ting Wang, Shuiping Huang, and Feng Chen. "Likelihood ratio tests in rare variant detection for continuous phenotypes." *Annals of human genetics* 78, no. 5 (2014): 320-332. https://doi.org/10.1111/ahg.12071

Zheng, Yan, Sylvia H. Ley, and Frank B. Hu. "Global aetiology and epidemiology of type 2 diabetes mellitus and its complications." *Nature reviews endocrinology* 14, no. 2 (2018): 88-98. https://doi.org/10.1038/nrendo.2017.151