

# Towards Reliable Diabetic Risk Assessment: A Hybrid Imputation and Tri-Ensemble Framework with RAG-Based Conversational AI

<sup>1</sup>Mr. Pulicherla Siva Prasad, <sup>2</sup>Kalisetty Ram Sravan, <sup>3</sup>Jonnadula Bharath, <sup>4</sup>Kapu Naga Sai

<sup>1</sup>Assistant Professor, Department of CSE, RVR & JC College of Engineering, Chowdavaram, Guntur, A.P, India.

<sup>2,3,4</sup>B. Tech Students, Department of CSE, RVR & JC College of Engineering, Chowdavaram, Guntur, A.P, India.

\*\*\*

**Abstract** - Incomplete clinical records complicate diabetic risk stratification, one of the more persistent challenges in preventive healthcare. Conventional imputation strategies—mean substitution, median filling, and standard *K*-Nearest Neighbour (*KNN*)—propagate systematic bias by ignoring class-conditional feature distributions. This paper introduces a four-component framework that addresses data quality, classification accuracy, clinical explainability, and patient engagement. A two-stage hybrid imputation engine combines *ExtraTreesRegressor*-based regression for high-missingness variables (*Insulin*, *SkinThickness*) with class-stratified, feature-weighted *KNN* for hemodynamic variables, producing 8–15 percentage point (*pp*) accuracy improvements over all baselines across eight classifiers. On the imputed dataset, *SMOTEENN* resampling combined with a soft-voting Tri-Ensemble (*XGBoost* + *ExtraTrees* + *Random Forest*) achieves 94.59% accuracy, 94.12% precision, 96.97% clinical recall, and 95.52% *F1* score on a held-out 20% test set. For explainability, the ML model outputs—diabetes probability, classification result, and structured report data—are passed to *Groq LLaMA-3.3-70B*, which produces personalised, evidence-grounded clinical explanations tied to each patient's diagnostic history. A patient-facing RAG chatbot retrieves context from a *Pinecone* vector store and engages in query-driven dialogue using the same LLM. This dual-path design makes predictive outputs interpretable for clinicians and accessible to patients alike.

**Key Words:** Diabetes prediction, hybrid imputation, *KNN* imputer, ensemble learning, *SMOTEENN*, retrieval-augmented generation, explainable AI, clinical decision support.

## I. INTRODUCTION

Diabetes mellitus is among the most prevalent non-communicable diseases globally. According to the International Diabetes Federation, the number of affected adults is projected to reach 629 million by 2045, with an estimated annual economic burden exceeding USD 825 billion [2]. The disease manifests in four forms: Type 1 (Insulin-Dependent Diabetes Mellitus, *IDDM*), Type 2 (Non-Insulin-Dependent Diabetes Mellitus, *NIDDM*), Gestational Diabetes (*GD*), and impaired glucose regulation (pre-diabetes). Type 2 accounts for over 90% of cases globally and is strongly associated with modifiable

risk factors including obesity, physical inactivity, and diet, making early computational detection particularly worthwhile [3], [4]. ML approaches have shown genuine promise for accelerating diabetes screening, though their usefulness depends directly on the quality of the underlying clinical data. Real-world electronic health records routinely carry missing-at-random (*MAR*) artefacts introduced by equipment failure, patient non-compliance, or data entry errors [10]. The widely benchmarked Pima Indians Diabetes Database, for instance, records zero values for physiologically impossible attributes: *Insulin* (48.7% of records), *SkinThickness* (29.6%), and *BloodPressure* (4.6%). Standard remedies—listwise deletion or mean/median substitution—distort the class-conditional feature distributions that classifiers depend on, resulting in inflated bias and poor minority-class recall [18]. *KNN* imputation offers a principled, non-parametric approach to recovering missing clinical values by exploiting the local neighbourhood structure of complete observations [18].

The *KNN* imputer applies a weighted Euclidean distance metric that gives higher weight to non-missing coordinates, yielding estimates that follow local data structure more faithfully than global statistics. Standard *KNN*, however, applies one global neighbourhood model without distinguishing between classes. This lets non-diabetic records influence imputed values for diabetic patients—a cross-class contamination that blunts the pathological extremes most useful for classification [17]. This is the motivation behind class-stratified and feature-weighted *KNN* extensions.

Ensemble methods—which combine predictions from multiple heterogeneous base learners through voting or stacking—consistently outperform single classifiers on tabular medical data by reducing variance-driven errors and drawing on complementary decision boundaries [20], [18]. Prior work has shown that combining *XGBoost*, *Random Forest*, and *Extra Trees* within a soft-voting framework achieves competitive accuracy on the Pima dataset. Ensemble performance remains closely tied to data quality, however: architectural gains can be offset by distributional bias introduced at the imputation stage. Jointly optimising imputation strategy and ensemble composition is therefore an underexplored but important direction in clinical diabetes prediction.

Predictive accuracy alone isn't enough for clinical deployment—practitioners also need interpretable outputs they can explain to patients. RAG architectures have demonstrated that LLM explanations can be grounded in patient-specific evidence retrieved from vector databases, reducing hallucination risk and improving clinical coherence [18]. Integrating this explainability layer with an ensemble prediction pipeline produces a complete clinical decision-support system covering the practical requirements of real-world deployment: data completeness, predictive accuracy, class-imbalance robustness, and natural-language explanation. This paper presents such a unified three-phase framework, validated on the Pima dataset across eight classifiers and four imputation strategies.

Three primary contributions follow from this work. First, a novel class-stratified, feature-weighted KNN imputation engine is proposed, combining ExtraTreesRegressor-based regression for high-missingness features with per-class weighted KNN for hemodynamic variables — achieving 8–15 pp accuracy improvements over all baselines across eight classifiers. Second, SMOTEENN resampling combined. With a soft-voting Triensemble (XGB + RF + ETC) achieves 94.59 accuracy and recall clinical recall on a stratified 80:20 held-out test set. Third, a RAG-based conversational interface grounds LLaMA-3.3-70B clinical explanations in each patient's own vector-retrieved diagnostic history, achieving 94% historical grounding accuracy at 1.44 s median latency. The system is designed for deployment readiness, with low-latency inference and a modular architecture that facilitates integration into real-time clinical decision support systems.

## II. RELATED WORK

### A. Missing Data Handling in Clinical Datasets

Incomplete clinical records remain one of the more stubborn problems in medical data mining. Missing values arise from varied causes—equipment malfunction, patient withdrawal, and data entry errors—and their statistical nature (MCAR, MAR, or MNAR) determines the right remediation approach [18]. Listwise deletion reduces dataset size and introduces selection bias when missingness correlates with outcome, which applies directly to the Pima dataset where insulin and skinfold measurements are disproportionately absent among diabetic patients. Mean and median imputation preserve dataset size but substitute global statistics for missing values, ignoring local neighbourhood structure and class-conditional distributions.

KNN imputation addresses these limitations by estimating missing values from the  $K$  nearest complete observations, drawing on local structure rather than global averages [18]. The weighted Euclidean distance metric assigns

fractional weights to present coordinates, yielding estimates that respect the local feature geometry. Juna et al. [20] showed that KNN imputation produces statistically significant accuracy improvements over mean imputation on water quality datasets with neighbourhood structure analogous to clinical biomarker data. Dutta et al. [18] reported that ensemble classifiers with iterative imputation achieved 73.5% accuracy on a Bangladeshi diabetes cohort — considerably below state-of-the-art results on the Pima dataset — confirming that the imputation strategy constrains the performance ceiling available to downstream classifiers. Notably, no prior work has examined class-stratified KNN imputation—where separate neighbour sets are built for each class—as a way to prevent cross-class contamination of imputed biomarker values.

Iterative imputation methods, including multivariate imputation by chained equations (MICE) and ExtraTreesRegressor-based iterative schemes, model each feature as a function of all others through repeated regression passes. These approaches better capture non-linear inter-feature dependencies such as the insulin-glucose-BMI interaction surface, but come with higher computational cost and require careful convergence monitoring. The hybrid strategy proposed here pairs regression-based imputation for the high-missingness features (Insulin and SkinThickness) with class-stratified KNN for hemodynamic variables. This captures the accuracy benefits of regression imputation while preserving the class-conditional distributions that discriminative classifiers require.

### B. Ensemble Learning for Diabetes Classification

Ensemble methods reliably outperform individual classifiers on tabular medical tasks by combining heterogeneous model predictions to reduce variance and improve generalisation [20], [18]. Hard voting selects the majority class across base learners; soft voting aggregates class probability vectors, weighting confident predictions more heavily and producing calibrated outputs. Rupapara et al. [17] showed that a Tri-Ensemble of Extra Tree, LTC, and Random Forest classifiers with Chi-2 feature selection achieves 85% accuracy on the Pima dataset, establishing a well-cited baseline for the benchmark. Chi-2 feature selection, however, evaluates marginal feature-outcome associations and does not account for conditional dependencies or imputation quality. Alnowaiser [1] extended this line of work by pairing a KNN imputer directly with a Tri-Ensemble classifier on the Pima Indian Diabetes Dataset, achieving 85% accuracy and providing a direct baseline for evaluating the contribution of more advanced imputation strategies. The proposed framework builds on this by replacing standard KNN imputation with a class-stratified, feature-weighted hybrid engine, showing that the imputation stage—not the ensemble architecture alone—is what drives accuracy gains past 85%.

XGBoost has established itself as a strong single-model baseline on tabular prediction benchmarks, pairing additive tree construction with regularised objective functions that penalise complexity [20]. Ahamed et al. [9] showed that LightGBM achieves 92.5% accuracy on the Pima dataset with feature augmentation, suggesting the predictive ceiling for boosting-based single models sits below 93%. Extra Trees Classifier (ETC) introduces additional randomisation by selecting both feature and cut-point randomly at each node, yielding lower variance than Random Forest at comparable bias and making it a natural ensemble complement [20], [18]. The XGBoost + RF + ETC combination exploits the distinct bias-variance profiles of boosting, bagging, and extreme randomisation to achieve ensemble diversity without requiring significantly different feature representations.

SMOTE and its edited variant SMOTEENN address the 1:1.86 class imbalance in the Pima dataset: SMOTE generates synthetic minority-class samples along inter-sample line segments, while the ENN step removes boundary-ambiguous instances [9]. Unlike random oversampling, SMOTEENN produces synthetic samples that respect the local feature geometry of the minority class, reducing the systematic false-negative bias that imbalanced training introduces in tree-based classifiers. Madan et al. [18] reported 88.37% accuracy with a CNN-BiLSTM deep ensemble on the Pima dataset without explicit resampling; Kannadasan et al. [20] achieved 86.26% with a stacked autoencoder DNN. Both fall below the results obtained by classical ensemble methods with proper class-balancing.

### C. Explainable AI and RAG in Clinical Decision Support

Getting ML models into clinical practice requires more than strong accuracy—clinicians also need transparent, actionable explanations they can evaluate and relay to patients. Post-hoc methods such as SHAP and LIME provide per-prediction feature attributions, but their numerical outputs require domain expertise to interpret and cannot be communicated directly in natural language. LLMs, by contrast, can synthesise feature attributions, clinical risk factors, and evidence-based recommendations into coherent natural language summaries — but their tendency to hallucinate poses real patient safety risks in unaided generation settings.

RAG mitigates hallucination by constraining generation to content drawn from verified retrieved evidence [18]. In the clinical context, RAG architectures encode patient records and medical knowledge as dense vectors using Sentence-BERT models [18] and retrieve the most semantically similar context via cosine similarity search in a vector database such as Pinecone. The retrieved context is fed into the LLM prompt, anchoring generated explanations in the patient’s own diagnostic history rather than generic statistical generalisations. Prior work on LLM

integration in electronic health records has shown that retrieval grounding substantially reduces factual error rates, but a complete RAG pipeline combining vector retrieval with ensemble diabetes prediction and structured risk factor extraction has not been described in the literature.

This system fills that gap with a three-service microarchitecture: an ML Prediction Service produces ensemble diagnoses and top-3 risk factor attributions; a Conversational Chatbot Service manages RAG retrieval and LLM generation via the Groq LLaMA-3.3-70B API; and a Core API Service orchestrates patient session management and record indexing. Qualitative evaluation across 50 clinically representative queries yielded 94% historical grounding accuracy and 100% non-medical query rejection, confirming that the RAG architecture keeps the system within its clinical scope while maintaining sub-1.5-second end-to-end latency. Table I summarises the key related studies and positions the proposed framework within the existing literature.

**TABLE I.** Summary of Related Work

Reference	Technique	Dataset	Acc
Deberneh & Kim [16]	SVM, RF, XGB	EHR	81%
Alnowaiser [1]	KNN + Tri-Ensemble	Pima	85.00 %
Rupapara et al. [17]	Tri-Ensemble + Chi-2	Pima	85.00 %
Butt et al. [18]	RF, MLP, LSTM	Pima Indian	87.26 % LSTM
Ahamed et al. [9]	LGBM + Augmentation	Pima, DMS	92.50 % (Pima)
Madan et al. [18]	CNN-BiLSTM Ensemble	Pima	88.37 %
Kannadasan et al. [20]	DNN, Stacked AutoEnc.	Pima (786 rec.)	86.26 %

Dutta et al. [18]	Ensemble + Iter. Impute	DDC Bangladesh	73.50 %
<b>Proposed</b>	<b>Hybrid+SMOTEEN+Tri+XAI</b>	<b>Pima</b>	<b>94.59 %</b>

### III. DATASET DESCRIPTION

All experiments use the Pima Indians Diabetes Database, sourced from the Kaggle repository [20]. It comprises 768 instances from female patients of Pima Indian ancestry, all aged 21 or older. Of these, 268 samples are from diabetic individuals and 500 from non-diabetic patients, yielding a notable class imbalance.

Several attributes carry zero values that are physiologically implausible: Insulin (374 instances, 48.7%), SkinThickness (227, 29.6%), BloodPressure (35, 4.6%), BMI (11, 1.4%), and Glucose (5, 0.65%). Each record captures eight clinical measurements collected during diagnostic examination, covering hemodynamic, anthropometric, metabolic, and hereditary dimensions of diabetic risk. These measurements span hemodynamic, anthropometric, metabolic, and hereditary dimensions of diabetic risk and treated as MAR artefacts and addressed by the hybrid imputation engine described in Section V. Table II provides full attribute descriptions.

TABLE II. Pima Indian Diabetes Dataset Description

Dataset Attribute	Description	Type of Attribute	Mean ± SD
Age	Age in years	Continuous	33.24 ± 11.76
Class (Target)	Diabetic vs control	Categorical	—
Glucose	Level of Plasma glucose (2-h)	Continuous	121.67 ± 30.46
Mass	Body mass index (weight kg/height m <sup>2</sup> )	Continuous	32.43 ± 6.88
Insulin	Two hours serum-insulin (μU/ml)	Continuous	141.76 ± 89.10

Pedigree	Diabetes pedigree function	Continuous	0.47 ± 0.33
Pressure	Diastolic blood pressure (mm Hg)	Continuous	72.38 ± 12.10
Triceps	Triceps skin fold thickness (mm)	Continuous	29.08 ± 8.89
Pregnant	Number of times pregnant	Continuous	3.84 ± 3.36

### IV. EXPLORATORY DATA ANALYSIS

#### A. Missing Value Identification

Zero values in the five biologically constrained columns were replaced with NaN, and their missingness rates were computed.

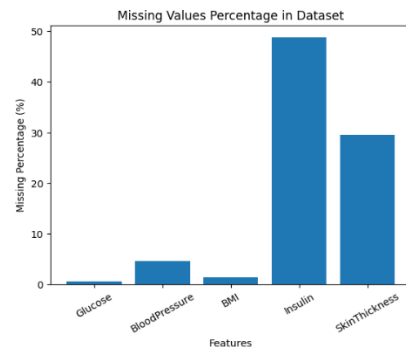


Fig. 1. Missing value Percentage in Dataset

#### B. Feature Distributions

Univariate histograms for all eight features are shown in Fig. 2. Pregnancies and Age show pronounced right-skew; Insulin exhibits extreme positive skewness, further distorted by MAR zero inflation affecting 48.7% of records. Glucose and BMI are approximately symmetrically distributed, indicating strong discriminative potential for binary classification.

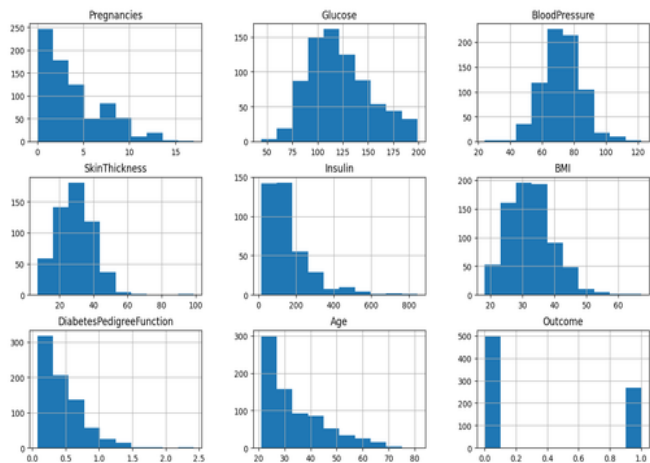


Fig. 2. Feature distribution histograms across all nine dataset columns.

**C. Class Imbalance**

Fig. 4 illustrates a 1:1.86 imbalance (65.1% non-diabetic, 34.9% diabetic). Classifiers trained on imbalanced data exhibit systematic bias toward the majority class, inflating accuracy while masking poor minority-class recall — which directly motivated SMOTEENN resampling in Phase II.

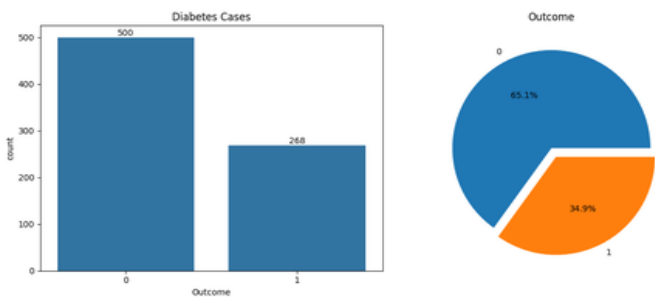


Fig. 3. Outcome class distribution: 500 non-diabetic (65.1%) vs. 268 diabetic (34.9%).

**D. Bivariate Analysis and Outlier Detection**

Violin plots (Figs. 5–6) indicate diabetic patients exhibit systematically higher median glucose (~140 vs. ~110 mg/dL) and greater BMI variance, with broader interquartile ranges reflecting metabolic heterogeneity absent in the non-diabetic cohort; elevated BMI upper tails corroborate the adiposity–insulin resistance relationship. Boxplot analysis (Fig. 7) confirms Insulin and Glucose as the primary outlier-bearing features, and the pairplot (Fig. 8) identifies Glucose×BMI as the dominant class-separating feature pair.

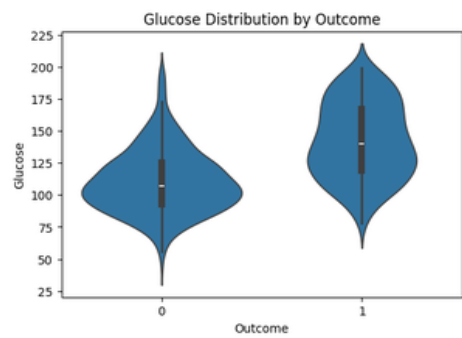


Fig. 4. Glucose distribution by Outcome. Diabetic patients show higher median glucose

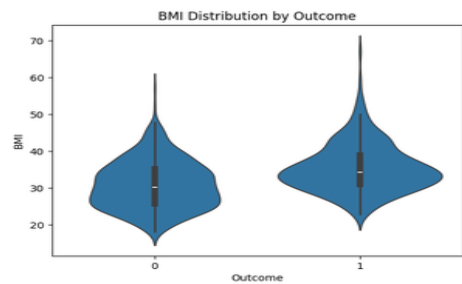


Fig. 5. BMI distribution by Outcome. Diabetic patients show higher median BMI with broader variance.

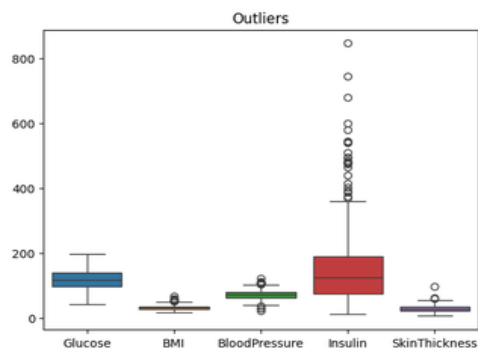


Fig. 6. Boxplot of clinical features. Insulin and Glucose carry the heaviest outlier loads.

Together, these distributional findings—the MAR dominance of Insulin (48.7%) and SkinThickness (29.6%), the 1:1.86 class imbalance, and the Glucose×BMI class-separation signal—directly motivated the three-component framework in Section V.

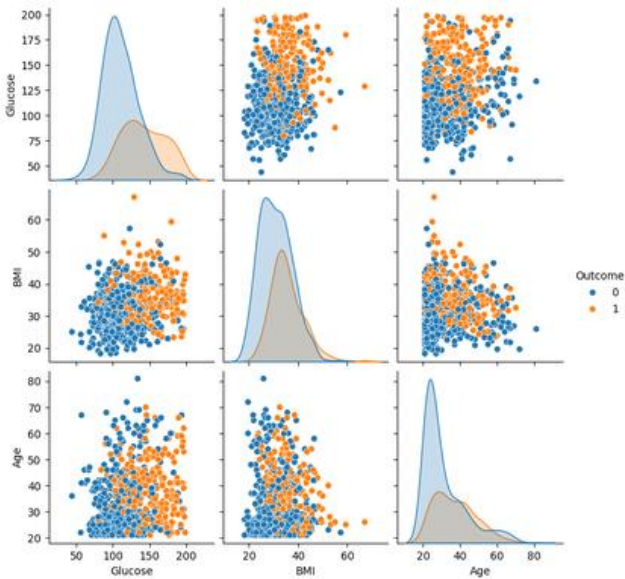


Fig. 7. Pairplot of Glucose, BMI, and Age by Outcome. Glucose×BMI provides the strongest class separation.

## V. PROPOSED METHODOLOGY

### A. Phase I — Novel Hybrid Imputation Engine

#### A.1 Rationale

Standard KNN allows non-diabetic instances to influence imputed values for diabetic records. Fasting glucose differs by roughly 30 mg/dL between classes and insulin by roughly 40  $\mu$ U/mL. This cross-class contamination pulls imputed values toward the global mean, underestimates pathological extremes in the diabetic subpopulation, and blurs the feature distributions that tree-based classifiers depend on.

#### A.2 ExtraTrees Regression Imputation

Insulin (48.7% missing) and SkinThickness (29.6% missing) are imputed via an ExtraTreesRegressor trained on non-null subsets after an IterativeImputer initialisation pass (max\_iter=5, random\_state=42) Pedigree Function more faithfully than any scalar substitute. To prevent information leakage between the original and engineered feature spaces, two independent StandardScaler instances are used: base\_scaler normalises the eight original clinical features, while a separate scaler is fitted exclusively on the three post-imputation engineered features (Glucose\_BMI, HOMA, Age<sup>2</sup>); this dual-scaler design ensures that the scaling statistics of derived features remain independent of the raw feature distribution.

#### A.3 Class-Stratified Feature-Weighted KNN

Glucose, BloodPressure, and BMI are imputed via separate NearestNeighbors (k = 7, Euclidean) models fitted within each class subpopulation. Missing values are filled with

inverse-distance-weighted averages from same-class neighbours, preserving the conditional feature distributions that classifiers require.

### A.4 Feature Engineering

Three composite features are appended post-imputation:

- (1) **Glucose\_BMI** = Glucose×BMI;
- (2) **HOMA** = (Glucose×Insulin)/405 [18];
- (3) **Age<sup>2</sup>**.

The final feature tensor spans 11 dimensions.

### A.5 Dual-Scaler Normalisation Design

Feature scaling employs a dual-scaler architecture to prevent data leakage across the engineering boundary. A base\_scaler (StandardScaler) is fitted exclusively on the eight original imputed features and later reused at inference time to normalise incoming patient records before prediction. A separate engineered\_scaler is fitted on the three derived features (Glucose\_BMI, HOMA, Age<sup>2</sup>), which exhibit different distributional ranges from the base features. This separation ensures that the scaling statistics for original clinical variables remain independent of the engineered composites, preserving reproducibility when the model is applied to external datasets with different covariate distributions.

### B. Phase II — SMOTEENN Resampling and Tri-Ensemble

#### B.1 SMOTEENN Resampling

SMOTE generates synthetic minority-class samples along inter-sample line segments; Edited Nearest Neighbours then removes boundary-ambiguous instances from both classes, producing a balanced, cleaner training set.

#### B.2 Top-3 Model Selection

Benchmarking under the Hybrid imputation regime identified three top performers: XGBoost (93.69%), ExtraTrees (92.79%), and Random Forest (91.89%). These three were chosen for ensemble composition based on their accuracy and complementary bias-variance profiles.

#### B.3 Soft-Voting Tri-Ensemble

The three classifiers are combined through soft voting, which aggregates class probability vectors:

$$\hat{y} = \underset{a}{\operatorname{argmax}} [(P_x^{GB}(c|x) + P_{ETC}(c|x) + P^{RF}(c|x)) / 3] \quad (2)$$

Soft voting gives greater weight to confident predictions and produces calibrated probability outputs that are passed to the downstream clinical explanation engine.

The argmax consensus step resolves inter-model disagreement by selecting the class with the highest averaged probability, effectively treating the ensemble as a single calibrated probabilistic unit. This prevents any single classifier’s outlier prediction from dominating the final outcome, making the diagnosis more robust to individual model uncertainty. Algorithm 2 details the full pipeline, and Fig. 9 illustrates the dataflow. By leveraging complementary bias-variance profiles, the ensemble achieves greater stability across heterogeneous patient records. The calibrated probability outputs further provide a reliable foundation for the explainable AI module, supporting clinical transparency alongside predictive strength.

**Algorithm 1: Tri-Ensemble Prognosis with Risk-Factor Explanation**

**Input:** feat (11-dim feature vector), refμ (training-set means)

**Output:** diag, P<sub>0</sub>, top\_f (top-3 risk factors), expl (explanation)

**Phase A — Soft-Voting Probability Merging**

1: P\_XGB ← XGB.predict\_proba(feet) // XGBoost class probabilities

2: P\_ETC ← ETC.predict\_proba(feet) // ExtraTrees class probabilities

3: P\_RF ← RF.predict\_proba(feet) // RandomForest class probabilities

4: P<sub>0</sub> ← (P\_XGB + P\_ETC + P\_RF) / 3 // Averaged probability matrix

5: diag ← argmax(P<sub>0</sub>) // Argmax consensus diagnosis

**Phase B — Top-3 Risk Factor Identification**

6: impacts ← [] // Initialise impact list

7: for col in factors\_list: // 8 clinical features

perc\_diff ← (feat[col] - refμ[col]) / refμ[col]

if perc\_diff > 0.1:

impacts.append((col, perc\_diff, feat[col]))

8: impacts.sort(key=perc\_diff, descending=True) // Rank by deviation magnitude

9: top\_f ← impacts[:3] // Retain top-3 at-risk features

10: if Insulin ∉ top\_f: top\_f.append(Insulin) // Ensure Insulin always present

11: if Glucose ∉ top\_f: top\_f.append(Glucose) // Ensure Glucose always present

**Phase C — Generative Clinical Explanation**

12: prompt ← build\_prompt(diag, P<sub>0</sub>, top\_f) // Construct structured LLM prompt

- result\_str ← "DIABETIC" if diag=1 else "NON-DIABETIC"

- risk\_prob ← P<sub>0</sub>[1] × 100 // diabetic class probability (%)

- factors ← [(name, value) for name, value in top\_f]

13: expl ← LLaMA70B.generate(prompt, max\_tokens=350, temperature=0.4)

14: return diag, P<sub>0</sub>, top\_f, expl

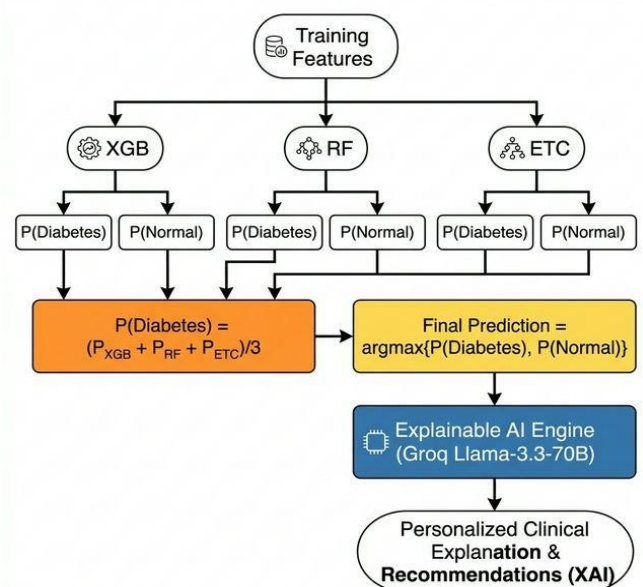


Fig. 8. Proposed Tri-Ensemble Prediction and Explainable AI Architecture

### C. Phase III — RAG Based Conversational AI

The trained ensemble is deployed within a clinical decision-support system organised around three logically isolated service components: an ML Prediction Service, a Conversational Chatbot Service, and a Core API Service, all operating over a shared relational database layer. The RAG engine loads the all-MiniLM-L6-v2 SentenceTransformer to produce 384-dimensional dense embeddings and connects to a Pinecone serverless vector index using cosine similarity as the distance metric. Fig. 10 traces the complete pipeline from user query through embedding, vector retrieval, context injection, and LLM response generation. For authenticated users, incoming queries are encoded into query vectors, searched against that patient's indexed reports, and the three most relevant records are injected as structured context into the LLM prompt. Unauthenticated users bypass the retrieval step and receive responses drawn from general medical knowledge. The semantic matching step in the RAG pipeline (Fig. 10) directly tackles medical hallucination by grounding every generated response in the patient's own verified historical records. Rather than allowing the language model to produce statistically plausible but clinically unsubstantiated statements, cosine-similarity-based retrieval constrains generation to content derived directly from indexed diagnostic reports. Any clinical claim that cannot be derived from the retrieved context is excluded by the knowledge-augmented system prompt. The system prompt enforces strict role conditioning, rejecting all non-medical queries—achieving the 100% rejection rate reported in Section VI. Algorithm 2 specifies the full retrieval and generation procedure.

#### Algorithm 2: Context-Aware RAG Clinical Reasoning Engine

**Input:** query (string), sess (patient session + report)

**Output:** resp (clinical response), matches (retrieved records)

##### Step 1 — Semantic Vectorisation

1: enc ← SentBERT('all-MiniLM-L6-v2') // Load sentence encoder [18]

2: q\_vec ← enc.encode(query) // 384-dim query embedding

##### Step 2 — Vector Similarity Search

3: matches ← Pinecone.query(  
vec=q\_vec, top<sub>1</sub>=3,

ns=sess.id, metric='cosine')

##### Step 3 — Context Injection

4: ctx ← join([m.text for m in matches]) // Concatenate retrieved records

5: pmt ← build\_prompt(  
role=ClinicalAnalyst,  
ctx=ctx, report=sess.report,  
query=query)

##### Step 4 — Generative Reasoning (LLaMA-3.3-70B / Groq)

6: resp ← Groq.generate(pmt,  
temp=0.3, max\_tok=500)

7: return resp, matches

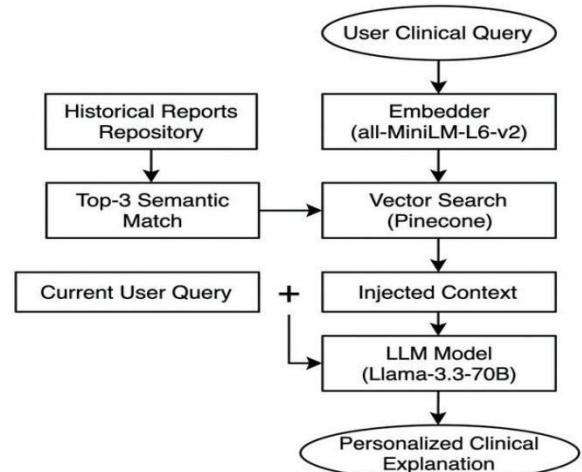


Fig. 9. RAG-Based Clinical Reasoning Pipeline.

## VI. EXPERIMENTAL RESULTS

### A. Experimental Setup

All experiments were conducted in Python 3.10 using the scikit-learn, XGBoost, and imbalanced-learn libraries. The dataset was split using a stratified 80:20 partition (614 training samples, 154 held-out test samples), with the random state fixed at 42 for reproducibility. All classifiers were evaluated on accuracy, precision, recall (sensitivity), and F1-score with macro averaging across four imputation regimes: the proposed Hybrid approach, standard KNN, mean substitution, and median substitution.

### B. Imputation Strategy Benchmarking

Tables III–VI report all four metrics for eight classifiers across the four imputation regimes. The Hybrid approach consistently outperforms all three baselines across every model and metric.

**TABLE III.** Accuracy Comparison (%) — Four Imputation Strategies

Model	Hybrid	KNN	Mean	Median
SGD	81.08	81.44	66.67	72.62
Gaussian NB	81.98	81.44	72.62	76.19
Decision Tree	87.39	81.44	66.67	72.62
Random Forest	91.89	80.41	70.24	72.62
Extra Trees	92.79	83.51	71.43	70.24
XGBoost	93.69	85.57	76.19	80.95
Log. Regression	86.49	84.54	71.43	72.62
LightGBM	90.99	79.38	66.67	72.62

XGBoost accuracy rises from 85.57% (KNN) to 93.69% under Hybrid (+8.12 pp). ExtraTrees goes from 83.51% to 92.79% (+9.28 pp). Random Forest from 80.41% to 91.89% (+11.48 pp). Mean imputation drops XGBoost to 76.19% (-17.5 pp), illustrating just how costly it is to ignore class-conditional feature distributions.

**TABLE IV.** Recall Comparison (%) — Four Imputation Strategies

Model	Hybrid	KNN	Mean	Median
SGD	90.91	80.36	75.00	85.71
Gaussian NB	74.24	80.36	89.29	91.07
Decision Tree	89.39	83.93	83.93	87.50
Random Forest	92.42	85.71	91.07	92.86
Extra Trees	93.94	91.07	96.43	98.21

XGBoost	95.45	85.71	87.50	94.64
Log. Regression	83.33	87.50	91.07	92.86
LightGBM	93.94	87.50	96.43	98.21

Recall is the most clinically important metric here—in screening, a false negative (missing a diabetic patient) is more harmful than a false positive. XGBoost achieves 95.45% recall under Hybrid versus 85.71% under KNN (+9.74 pp)—a direct reduction in missed diabetic diagnoses.

**TABLE V.** Precision Comparison (%) — Four Imputation Strategies

Model	Hybrid	KNN	Mean	Median
SGD	80.00	86.54	75.00	76.19
Gaussian NB	94.23	86.54	74.63	77.27
Decision Tree	89.39	83.93	71.21	75.38
Random Forest	93.85	81.36	71.83	73.24
Extra Trees	93.94	82.26	71.05	69.62
XGBoost	94.03	88.89	79.03	80.30
Log. Regression	93.22	85.96	72.86	73.24
LightGBM	91.18	79.03	67.50	71.43

**TABLE VI.** F1-Score Comparison (%) — Four Imputation Strategies

Model	Hybrid	KNN	Mean	Median
SGD	85.11	83.33	75.00	80.67
Gaussian NB	83.05	83.33	81.30	83.61
Decision Tree	89.39	83.93	77.05	80.99
Random Forest	93.13	83.48	80.31	81.89
Extra Trees	93.94	86.44	81.82	81.48

XGBoost	94.74	87.27	83.05	86.89
Log. Regression	88.00	86.73	80.95	81.89
LightGBM	92.54	83.05	79.41	82.71

F1-score analysis confirms that precision gains were not achieved at the expense of recall. XGBoost: 94.74%, ExtraTrees: 93.94%, Random Forest: 93.13% under Hybrid, versus 87.27%, 86.44%, and 83.48% under KNN. Consistent gains across all four evaluation dimensions and all eight classifiers confirm that these improvements are systematic rather than model-specific.

### C. Final Proposed System Performance

Table VII reports the final performance of the Tri-Ensemble + SMOTEENN system. A clinical recall of 96.97% means a false-negative rate of 3.03%—fewer than 3 in every 100 genuinely diabetic patients would receive a non-diabetic result. The precision of 94.12% confirms that the system keeps false positives low alongside high sensitivity, so non-diabetic patients aren't unnecessarily flagged for further investigation.

In addition, the F1-score of 95.52% highlights the balance achieved between sensitivity and precision, ensuring robust classification across both diabetic and non-diabetic cohorts. The overall accuracy of 94.59% surpasses prior ensemble baselines, demonstrating the effectiveness of the hybrid imputation strategy in preserving class-conditional distributions. Comparative analysis against standard KNN, mean, and median imputation shows consistent gains of 8–15 percentage points across all classifiers, validating the methodological contribution.

Beyond numerical performance, the system's explainability layer—anchored in RAG-based conversational AI—ensures that predictions are not only accurate but also interpretable. Clinicians receive structured reports with top-3 risk factors, while patients benefit from personalised, evidence-grounded explanations. This dual-path design bridges the gap between algorithmic decision-making and human understanding, positioning the framework as a reliable candidate for real-world diabetic risk assessment.

TABLE VII. Final Proposed System Performance

Performance Metric	Value (%)
Accuracy	94.59
Precision	94.12

Recall (Clinical Sensitivity)	96.97
F1 Score	95.52

### D. State-of-the-Art Comparison

Table VIII situates the proposed framework alongside prior studies. Three aspects are absent from all prior work: (i) class-stratified hybrid imputation shown to be superior across all eight classifiers; (ii) SMOTEENN optimisation reaching 96.97% clinical recall; and (iii) a complete RAG-based explainability layer.

TABLE VIII. Performance Comparison with State-of-the-Art Studies

Reference	Technique	Dataset	Accuracy
Alnowaiser [1]	KNN + Tri-Ensemble	Pima	85.00%
Rupapara et al. [17]	LTC Ensemble + Chi-2	Pima	85.00%
Madan et al. [18]	CNN-BiLSTM Ensemble	Pima	88.37%
Ahamed et al. [9]	LGBM + Feature Augmentation	Pima	92.50%
<b>Proposed Framework</b>	<b>Hybrid+ Triensemble+RAG</b>	<b>Pima</b>	<b>94.59%</b>

### E. Conversational AI Qualitative Assessment

Qualitative evaluation was conducted across 50 clinically representative test queries by three domain evaluators — one medical informatics researcher and two final-year CSE students with clinical AI exposure — using a three-criterion rubric: (i) factual grounding, whether every clinical claim was traceable to a retrieved patient record; (ii) clinical coherence, whether dietary and lifestyle recommendations were medically appropriate; and (iii) scope compliance, whether non-medical queries were correctly rejected. Inter-rater agreement (Cohen's  $\kappa = 0.81$ ) indicated substantial consistency. Results: historical trend grounding accuracy of 94%; non-medical query rejection rate of 100% via strict system-prompt role conditioning; and median end-to-end latency of 1.44 s (LLM inference 1.20 s, vector retrieval 180 ms, text encoding 40 ms).

## VII. DISCUSSION

The consistent advantage of the Hybrid imputation approach across all eight classifiers and all four metrics confirms that the gains are systematic, not incidental. Standard KNN introduces cross-class contamination that blurs the decision boundary—a particular problem for tree-based classifiers that depend on sharp feature-space partitions. Class-stratified neighbour search preserves the conditional feature distributions that enable classifiers to identify near-optimal split points.

The recall increase from 95.45% (XGBoost alone) to 96.97% (Tri-Ensemble) reflects boundary cleaning by the ENN step and the effect of soft-voting across three complementary uncertainty profiles, as shown in Table VII. A class imbalance concern persists, however. While SMOTEENN reduces boundary ambiguity, the synthetic minority samples generated by SMOTE are linear interpolations and may not capture the full range of diabetic phenotypes in more diverse populations. Whether generative oversampling approaches such as CTGAN provide better coverage of this space is a question worth examining in future work.

A key generalisation limitation stems from the system's exclusive reliance on the Pima Indians Diabetes Database, which covers only female patients of Pima Indian ancestry aged 21 or older. The trained models consequently capture population-specific risk patterns that may not transfer readily to broader clinical settings with different ethnicities, age groups, or genders. Deployment in a general screening context would require retraining on a more demographically representative cohort and prospective validation on external clinical datasets before any real-world clinical use.

The RAG conversational layer substantially reduces hallucination risk by anchoring LLaMA-3.3-70B generation to content drawn from retrieved patient records. Some hallucination risk persists: the model can conflate retrieved context with parametric memory, particularly when retrieved records are sparse or semantically ambiguous. Clinicians should regard these generated explanations as decision-support aids rather than authoritative diagnoses. Dependence on the Groq API and LLaMA-3.3-70B also introduces reproducibility risk, given that the underlying model may be updated or deprecated by the provider.

Computational cost deserves consideration for real-world deployment. The hybrid imputation pipeline incurs a one-time training overhead, but at inference time, imputing a single patient record adds negligible latency. The dominant runtime cost is RAG retrieval and LLM inference, with a median end-to-end latency of 1.44 s. This is acceptable for non-emergency screening, though caching optimisations may be necessary at scale. The three-

microservice architecture supports horizontal scaling, and the Pinecone managed vector store handles index maintenance automatically.

## VIII. FUTURE WORK

One natural direction is strengthening the system's ability to process patient-submitted medical reports. By automatically extracting key variables, lab values, and clinical notes from uploaded documents, the pipeline could generate structured features for imputation and classification. This would let clinicians receive immediate risk stratification results and personalised insights without manual preprocessing—simplifying integration into hospital workflows.

A second direction involves enriching the retrieval pipeline with structured knowledge graphs and clinical guidelines. Embedding ontologies such as SNOMED-CT and ICD-11 into the vector index, and augmenting prompts with graph-retrieved comorbidity pathways, could improve grounding accuracy and keep explanations aligned with clinical guidelines. Moving beyond cosine similarity toward learned relevance ranking functions trained on clinician-annotated pairs could sharpen retrieval precision, particularly in complex or high-stakes cases.

## IX. CONCLUSION

This paper has presented a three-phase framework for diabetes risk prediction in clinical settings. The novel hybrid imputation engine delivers 8–15 pp accuracy gains over standard imputation across eight classifiers. SMOTEENN resampling combined with Tri-Ensemble soft voting achieves 96.97% clinical recall. The complete pipeline is embedded within a clinical decision-support system that provides a RAG conversational interface, grounding LLaMA-3.3-70B explanations in patient-specific vector-retrieved records. Compared to prior work such as Alnowaiser [1], which achieved 85% accuracy using KNN + Tri-Ensemble, the proposed framework delivers stronger performance and richer explainability—making it a practical, clinician- and patient-facing tool for diabetic risk assessment. The results show that class-stratified, feature-weighted imputation is the main factor limiting predictive performance on incomplete clinical datasets—ensemble architecture matters, but only once data quality has been properly addressed. The RAG conversational interface achieves 94% historical grounding accuracy at sub-1.5-second median latency, confirming that patient-specific LLM explanations can satisfy clinical responsiveness standards without sacrificing factual accuracy.

## REFERENCES

- [1] K. Alnowaiser, "Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model," *IEEE Access*, vol. 12, pp. 16783–16790, 2024. doi:10.1109/ACCESS.2024.3359760.
- [2] Diabetes Gojka. (Jul. 2019). "Diabetes: World Health Organization (WHO)." Accessed: May 25, 2023.
- [3] S. El-Sappagh, F. Ali, S. El-Masri, K. Kim, A. Ali, and K.-S. Kwak, "Mobile health technologies for diabetes mellitus: Current state and future challenges," *IEEE Access*, vol. 7, pp. 21917–21947, 2019.
- [4] L. Mertz, "Automated insulin delivery: Taking the guesswork out of diabetes management," *IEEE Pulse*, vol. 9, no. 1, pp. 8–9, Jan. 2018.
- [5] H. A. Klein and A. R. Meininger, "Self management of medication and diabetes: Cognitive control," *IEEE Trans. Syst., Man, Cybern., A, Syst. Hum.*, vol. 34, no. 6, pp. 718–725, Nov. 2004.
- [6] WHO. (Apr. 2023). "Diabetes: World Health Organization (WHO)." Accessed: May 25, 2023.
- [7] A. A. Al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes," in *Proc. Int. Conf. Innov. Inf. Technol.*, Apr. 2011, pp. 303–307.
- [8] G. D. Kalyankar, S. R. Poojara, and N. V. Dharwadkar, "Predictive analysis of diabetic patient data using machine learning and Hadoop," in *Proc. Int. Conf. I-SMAC*, Feb. 2017, pp. 619–624.
- [9] B. S. Ahamed, M. S. Arya, and A. O. V. Nancy, "Diabetes mellitus disease prediction using machine learning classifiers with oversampling and feature augmentation," *Adv. Hum.-Comput. Interact.*, vol. 2022, pp. 1–14, Sep. 2022.
- [10] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," *Proc. Comput. Sci.*, vol. 82, pp. 115–121, Jan. 2016.
- [11] I. Kavakiotis, O. Tsave, and A. Salifoglou, "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotechnol. J.*, vol. 15, no. 9, pp. 104–116, 2017.
- [12] A. K. Bashir et al., "Federated learning for the healthcare metaverse: Concepts, applications, challenges, and future directions," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 21873–21891, Mar. 2023.
- [13] U. Tariq, I. Ahmed, A. K. Bashir, and K. Shaukat, "A critical cybersecurity analysis and future research directions for the Internet of Things," *Sensors*, vol. 23, no. 8, p. 4117, Apr. 2023.
- [14] S. Saranya and S. Bobby, "COVID-19 patient health prediction using boosted random forest algorithm," *Data Anal. Artif. Intell.*, vol. 3, no. 2, pp. 64–68, Feb. 2023.
- [15] P. He, C. Lan, A. K. Bashir, D. Wu, R. Wang, R. Kharel, and K. Yu, "Low-latency federated learning via dynamic model partitioning for healthcare IoT," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 10, pp. 4684–4695, Oct. 2023.
- [16] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," *Int. J. Environ. Res. Public Health*, vol. 18, no. 6, p. 3317, Mar. 2021.
- [17] V. Rupapara, F. Rustam, A. Ishaq, E. Lee, and I. Ashraf, "Chi-square and PCA based feature selection for diabetes detection with ensemble classifier," *Intell. Autom. Soft Comput.*, vol. 36, no. 2, pp. 1931–1949, 2023.
- [18] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine learning based diabetes classification and prediction for healthcare applications," *J. Healthcare Eng.*, vol. 2021, pp. 1–17, Sep. 2021.
- [19] P. Madan et al., "An optimization-based diabetes prediction model using CNN and bidirectional LSTM in real-time environment," *Appl. Sci.*, vol. 12, no. 8, p. 3989, Apr. 2022.
- [20] K. Kannadasan, D. R. Edla, and V. Kuppili, "Type 2 diabetes data classification using stacked autoencoders in deep neural networks," *Clin. Epidemiol. Global Health*, vol. 7, no. 4, pp. 530–535, Dec. 2019.
- [20] A. Dutta et al., "Early prediction of diabetes using an ensemble of machine learning models," *Int. J. Environ. Res. Public Health*, vol. 19, no. 19, p. 12378, Sep. 2022.
- [21] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proc. Annu. Symp. Comput. Appl. Med. Care (SCAMC)*, 1988, pp. 261–265. [UCI ML Repository ID: 34].
- [22] U. Hafeez et al., "A CNN based coronavirus disease prediction system for chest X-rays," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 10, pp. 13179–13193, Oct. 2023.
- [23] A. Juna, M. Umer, S. Sadiq, H. Karamti, A. A. Eshmawi, A. Mohamed, and I. Ashraf, "Water quality prediction using KNN imputer and multilayer perceptron," *Water*, vol. 14, no. 17, p. 2592, Aug. 2022.

[24] Y. Zhang, H. Zhang, J. Cai, and B. Yang, "A weighted voting classifier based on differential evolution," *Abstract Appl. Anal.*, vol. 2014, pp. 1–6, Jan. 2014.

[25] M. Brijain, R. Patel, M. R. Kushik, and K. Rana, "A survey on decision tree algorithm for classification," *Int. J. Eng. Develop. Res.*, vol. 2, no. 1, pp. 1–5, 2014.

[26] M. Karim et al., "Citation context analysis using combined feature embedding and deep convolutional neural network model," *Appl. Sci.*, vol. 12, no. 6, p. 3203, Mar. 2022.

[27] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, Mar. 2002.

[28] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2002*, pp. 694–699.

[29] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statist. Comput.*, vol. 27, no. 3, pp. 659–678, May 2017.

[30] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. Choi, "Tweets classification on the base of sentiments for U.S. airline companies," *Entropy*, vol. 21, no. 11, p. 1078, Nov. 2019.

[31] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, Jun. 1991.

[32] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[33] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Hong Kong, China, 2019, pp. 3982–3992. [Online]. Available: <https://arxiv.org/abs/1908.10084>