THMI-FS-Stack: A Hybrid Imputation and Feature Selection with Stacking Ensemble for Avian Influenza Outbreak Prediction

V. S. V. S. Murthy[1], J. N. V. R. Swarup Kumar[2], Srinivas Gorla[3]
Research Scholar, Dept. of CSE, GST, GITAM Deemed to be University, Visakhapatnam, India[1]
Dept. of CSE, GST, GITAM Deemed to be University, Visakhapatnam Campus, India[2]
Professor and Dean, Dept. of CSE, ANITS, Visakhapatnam, India[3]

***Abstract*—Timely prediction of zoonotic disease outbreaks, particularly Highly Pathogenic Avian Influenza (HPAI), is critical for real-time epidemiological surveillance and pandemic preparedness. However, real-world avian surveillance datasets often suffer from missing values, high dimensionality, and inconsistent feature distributions, leading to unreliable predictions. This study proposes THMI-FS-Stack, a modular and interpretable machine learning pipeline that integrates hybrid data imputation, scalable feature selection, and ensemble classification for outbreak forecasting. The first stage, THMI-CB, employs a two-layer imputation framework combining statistical techniques (Mode, Hot Deck, KNN) and machine learning models (Bayesian Networks and CatBoost), achieving an F1-score of 0.91. The second stage, Hybrid-FS-ML, combines filter-based ranking (Mutual Information, Chi-Square, mRMR) with wrapper-based optimization using a Genetic Algorithm, achieving a 72% dimensionality reduction and an F1-score of 0.96. The final component is a stacking ensemble classifier that uses Random Forest and XGBoost as base learners and Logistic Regression as the meta-learner, yielding an F1-score of 0.92, accuracy of 0.93, and AUC-PR of 0.89. Evaluated on the Wild Bird HPAI dataset with 5-fold stratified cross-validation, THMI-FS-Stack consistently outperforms baseline models. Its robust architecture, low computational cost (runtime of 28–36s), and strong generalization ability make it highly suitable for noisy, incomplete epidemiological data in wildlife surveillance dashboards and early-warning systems.**

***Keywords*—Zoonotic disease outbreaks; avian influenza; pandemic prediction; hybrid imputation; feature selection; stacking ensemble classifier; wild bird HPAI dataset; early-warning systems**

I. Introduction

Avian Influenza (AI), especially the highly pathogenic subtypes like H5N1 and H7N9, presents a significant threat to global public health and biodiversity. AI outbreaks in wild birds often serve as early indicators for potential zoonotic transmissions, making timely surveillance and accurate outbreak prediction essential. Delayed detection not only hampers containment but also causes widespread socio-economic impacts in the poultry industry and public health sectors [1]. Predictive modeling, powered by ecological and epidemiological data, has emerged as a valuable tool for early warning systems. However, building reliable models in this domain is hindered by two key challenges: missing data and high-dimensional feature spaces [2], [3].

The Wild Bird HPAI dataset, used in this study, is an open-source resource that reflects real-world complexities such as data incompleteness due to irregular monitoring, delayed reporting, and inconsistent entries. Additionally, the dataset contains a large number of features, many of which are redundant or weakly informative, contributing to overfitting and reduced model interpretability.

To address the missing data problem, traditional imputation methods like Mode, Hot Deck, and KNN are widely used for their simplicity, yet they often fail to capture nonlinear dependencies or contextual patterns. Recent studies explore advanced machine learning-based imputations, including Bayesian Networks and CatBoost [4], which improve accuracy by modeling inter-feature relationships. Deep learning methods such as GAIN and Transformers further advance imputation performance but are often limited by computational cost and lack of transparency.

Similarly, in high-dimensional settings, selecting relevant features is critical for improving generalization. Filter-based approaches (e.g., Mutual Information, Chi-Square, mRMR) are fast but univariate. Wrapper methods like Genetic Algorithms offer superior optimization but are computationally expensive. A hybrid approach that combines both strategies provides a practical balance.

To overcome these challenges, we propose THMI-FS-Stack, a unified machine learning framework integrating three modules: (1) HMI-CB, a two-layer hybrid imputation strategy combining statistical and machine learning methods, (2) Hybrid-FS-ML, a two-phase feature selection pipeline that leverages filter-based ranking and wrapper-based Genetic Algorithm optimization, and (3) a Stacking Ensemble Classifier combining Random Forest, XGBoost, and Logistic Regression to produce robust outbreak predictions.

Fig. 1 illustrates the end-to-end pipeline. The framework is modular, interpretable, and designed for scalability in real-time surveillance settings. It is validated on the Wild Bird HPAI dataset using 5-fold stratified cross-validation.

This study offers the following novel, quantifiable, and domain-specific contributions:

- THMI-FS-Stack pipeline: A unified framework integrating imputation, feature selection, and classification, unlike prior works treating them as separate tasks.
- Hybrid imputation (THMI-CB): Combines statisti-

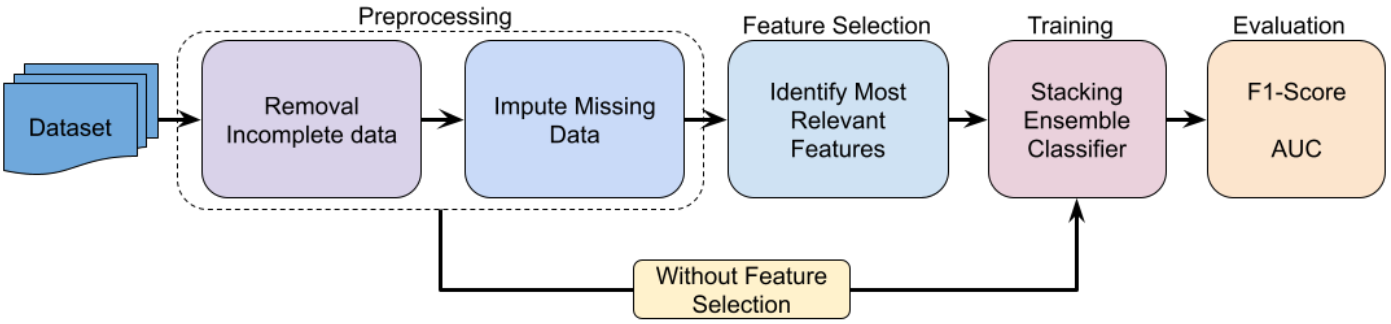

Fig. 1. Workflow of the proposed THMI-FS-Stack framework for Avian Influenza outbreak prediction.

cal (Mode, Hot Deck, KNN) and machine learning (Bayesian Networks, CatBoost) methods, improving F1-score by 14% over traditional and 2–4% over standalone ML/DL baselines.

- Hybrid feature selection (Hybrid-FS-ML): Merges filter-based (MI, Chi-Square, mRMR) and wrapper-based (GA) methods, achieving 72% dimensionality reduction and up to 8% F1-score gain.
- Stacking classifier performance: Combines Random Forest, XGBoost, and Logistic Regression, achieving 5% higher F1-score and 4% better AUC-PR over AdaBoost and Bagging.
- First-of-its-kind for HPAI surveillance: Tailored for noisy, incomplete, and high-dimensional avian outbreak data, bridging methodological rigor with practical relevance.

The rest of the paper is organized as follows: Section II reviews related work on imputation, feature selection, and ensemble learning. Section III describes the dataset and preliminaries. Section IV presents the proposed THMI-FS-Stack methodology in detail. Section V discusses the experimental setup, evaluation metrics, and results. Section VI concludes the paper and outlines future directions.

II. Related Study

The prediction of Avian Influenza outbreaks is complex, mainly due to the challenges of missing data, high dimensionality, and the dynamic nature of the features involved [3]. Many studies have explored various data imputation techniques, feature selection methods, and ensemble learning models to overcome these challenges in the broader field of disease prediction. This section summarizes the key methodologies used in the literature to deal with missing data, optimize features, and improve predictive accuracy in epidemiological forecasting, with a specific focus on Avian Influenza outbreaks.

A. Missing Value Imputation Techniques

The handling of missing data [5] is a critical task in disease prediction models. This is because insufficient data can cause biased results or decrease model accuracy. Several methods have been proposed for imputing missing values in epidemiological datasets [2]. Traditional imputation methods like mean imputation, mode imputation, and K-Nearest Neighbors are extensively used due to their simplicity and computational efficiency [6], [3]. Mean and mode imputation are suitable for cases where data is “Missing Completely At Random”. But these methods may fail when data is “Missing at Random” or “Missing Not At Random”. KNN imputation uses feature similarity to impute missing values. It has performed better in certain models [6], [7].

Machine learning-based imputation techniques have been developed to handle epidemiological datasets that possess complex feature dependencies. For example, Bayesian Networks (BNs) can model conditional dependencies between features, making them well-suited for imputing missing data when the relationships among features are known [8]. CatBoost is a gradient-boosting algorithm explicitly developed for categorical data. It has been shown to impute missing values more accurately by learning feature interactions during model training [9].

Deep learning-based models like Generative Adversarial Imputation Networks use adversarial training to produce plausible missing data samples based on observed distributions [8]. These approaches are effective but computationally expensive and require large datasets for training. Conversely, traditional methods such as Hot Deck Imputation and Multiple Imputation remain practical for smaller datasets and environments with limited resources [10].

The THMI-CB hybrid imputation framework introduced in this paper seeks to combine the strengths of traditional methods (such as Mode, Hot Deck, and KNN) with machine learning-based techniques (such as CatBoost and Bayesian Networks) to handle both simple and complex missingness patterns efficiently.

B. Feature Selection Methods in Disease Prediction

Feature selection is essential for boosting model performance. This procedure is particularly necessary when processing high-dimensional datasets where irrelevant or redundant features reduce the accuracy of predictive models. In the context of Avian Influenza outbreak prediction, choosing the relevant features is essential for attaining accurate and interpretable predictions [11], [12]. A range of feature selection methods have been proposed, like, filter, wrapper, and embedded methods [13]. Filter-based methods evaluate the significance of features by analyzing their statistical relationship with

TABLE I. COMPARISON OF RELATED WORKS IN IMPUTATION, FS, AND CLASSIFICATION

Study	Methods and Key Outcomes
Alnowaiser et al. [3]	Mode, KNN + MI, Chi-Square + RF/SVM. Improved accuracy with hybrid FS.
Fan et al. [4]	BN + mRMR + XGBoost. ML-based imputation and FS improved predictions.
Xue et al. [11]	GAIN + GA + Stacking. Deep imputation enhanced AI forecasting.
Khan et al. [10]	Hot Deck + GA + SVM. Categorical imputation improved results.
Proposed (THMI-FS-Stack)	Hybrid (Stat+ML) Imputation + Hybrid FS (Filter+GA) + Stacking (RF+XGB+LR). Outperformed all baselines in F1/AUC.

the target variable [14]. Mutual Information (MI), Chi-Square, and mRMR (Minimum Redundancy Maximum Relevance) are the techniques used for ranking features based on their individual relevance and redundancy [9]. These methods are fast. But they do not consider feature dependencies, which can lead to the omission of important feature combinations.

Wrapper-based methods, such as Recursive Feature Elimination and Genetic Algorithms, are more sophisticated as they evaluate feature subsets by training models on them. GA-based feature selection methods iteratively search for optimal feature combinations by assessing the performance of a model trained on the selected features [10]. Genetic Algorithm-based feature selection can improve prediction accuracy by choosing relevant features that maximize model performance. Embedded methods, such as Lasso regression and Random Forest feature importance, integrate feature selection with model training. These methods are mainly effective when the classification model itself has built-in feature selection capabilities. For example, Random Forests can rank features by evaluating the impurity reduction caused by each feature during training [8].

The Hybrid-FS-ML framework proposed here combines filter-based methods such as MI, Chi-Square, and mRMR with wrapper-based Genetic Algorithm optimization to recognize the most informative and non-redundant features for the model. The next section presents the design and implementation of the THMI-FS-Stack framework, which integrates the above strategies into a unified pipeline. It uses the combined power of traditional and machine learning-based imputation, hybrid feature selection, and a stacking-based ensemble classifier to address the challenges of noisy, incomplete, and high-dimensional avian influenza surveillance data.

C. Ensemble Learning and Stacking for Disease Prediction

Ensemble learning methods that combine multiple models to enhance predictive accuracy [15] have become a standard approach in machine learning. Stacking is one of the most extensively used ensemble techniques, where predictions from base models are combined using a meta-model [16], [17]. Stacking has been successfully applied in a wide range of domains like healthcare and epidemiological forecasting, to enhance the robustness and generalization of predictive models [18], [19]. Multiple base models like Random Forests, XGBoost, and Support Vector Machines are trained on the data in the stacking approach. Their predictions are used as input to a meta-learner, typically Logistic Regression or Linear Regression. These techniques reduce overfitting. They also increase accuracy by capturing diverse decision boundaries from base models [20].

In the case of Avian Influenza outbreak prediction, stacking ensemble can enhance performance by combining the strengths of different classifiers. Random Forest and XGBoost are highly effective in handling non-linear relationships and feature interactions. Logistic Regression, being a meta-learner, can integrate the outputs of these models to improve predictive accuracy. Stacking ensemble have been utilized in disease forecasting, where it improved accuracy compared to individual models like SVM or Random Forests [8]. Table I summarizes the key studies in prediction based on machine learning and deep learning models and their relevance to the current work. The THMI-FS-Stack framework apply Random Forest and XGBoost as base classifiers and combines them with Logistic Regression as the meta-learner. In this way, the framework harnesses the power of stacking ensemble models. This allows the model to learn complex patterns in the data and enhance its predictive capability for Avian Influenza outbreaks.

D. Research Gap

Despite significant progress in the field of AI outbreak prediction, several gaps remain in current methodologies. First, most studies treat missing data imputation, feature selection, and classification as separate tasks, rather than integrating them into a unified framework. While machine learning-based imputation techniques, such as Bayesian Networks and CatBoost, have demonstrated their effectiveness, they are not always incorporated into hybrid models that combine traditional and advanced imputation methods. Moreover, ensemble methods, including stacking, have shown promise in improving predictive performance but are rarely combined with hybrid feature selection and imputation models for complex epidemiological datasets like the Wild Bird HPAI dataset. Secondly, there is a lack of studies comprehensively addressing the high-dimensionality problem in AI outbreak prediction. The vast number of spatiotemporal features present in the Wild Bird HPAI dataset makes it crucial to select only the most relevant features. The integration of filter-based and wrapper-based feature selection methods has not been fully explored in disease prediction models, especially for epidemiological data with missing values. Finally, while several studies have utilized stacking ensemble models, there is limited research that incorporates hybrid imputation, feature selection, and ensemble learning into a single framework for AI outbreak prediction.

III. MATERIALS AND METHODS

A. Dataset

This study utilizes the Wild Bird Highly Pathogenic Avian Influenza (HPAI) dataset [21], an authoritative and structured epidemiological resource designed for avian flu surveillance. The dataset, sourced from government and international monitoring programs, contains extensive records on wild bird species across diverse geographical locations and timeframes.

It encompasses both numerical and categorical features essential for modeling avian influenza outbreaks. Table II summarizes the core characteristics of the processed Wild Bird HPAI dataset.

TABLE II. SUMMARY OF WILD BIRD HPAI DATASET CHARACTERISTICS

Attribute	Description
Number of Records	8,351 (after cleaning)
Features	41 total (24 categorical, 17 numerical)
Missingness Pattern	14 features with MCAR/MAR, avg. 22.3% missing
Class Imbalance	Positive: 13.1% (HPAI), Negative: 86.9%
Temporal Distribution	Skewed toward 2021–2022 peak outbreaks
Spatial Coverage	70 countries across 6 continents

The dataset includes the following major categories of features:

- Bird Identification: For species-level analysis, scientific and common names (e.g., *Anas platyrhynchos*).
- Temporal Information: Timestamped details including date, year, month, day, and time of observation, facilitating seasonal trend analysis.
- Geospatial Attributes: Geographic indicators such as country, state, county, locality, latitude, and longitude, enabling spatial mapping of outbreak risks.
- Biological and Health Status: Taxonomic hierarchy (e.g., bird family, parent species) and disease outcome variable (HPAI positive/negative).

The dataset suffers from extensive missing values in critical categorical fields such as *County*, *Locality*, and *Bird Family*, with missingness rates exceeding 35% in some columns. Overall, more than 22% of instances contain at least one missing field. The nature of missingness varies: *Locality* and *County* follow Missing at Random (MAR) patterns linked to *Region* and *Season*, while fields like *Species* exhibit Missing Completely at Random (MCAR). Additionally, the dataset is imbalanced, with fewer than 15% of samples labeled as HPAI-positive. Temporal and spatial skew is also present: some southern states and winter seasons dominate the dataset, leading to biased distributions. These challenges, namely, sparsity, skew, and imbalance, necessitate robust hybrid imputation (THMI-CB), intelligent feature selection (Hybrid-FS-ML), and ensemble modeling strategies to mitigate bias and improve prediction accuracy.

B. Methods

The proposed THMI-FS-Stack framework addresses three major challenges in Avian Influenza outbreak prediction: (i)missing data,(ii) high-dimensional feature spaces, and (iii) predictive robustness. The methodology integrates three primary stages: (i) imputation,(ii) feature selection, and (iii) classification. Each stage is designed using a combination of traditional, machine learning, and deep learning techniques. This section details the rationale, techniques, and mathematical formulations employed at each stage.

1) Imputation methods: A hybrid imputation strategy (THMI-CB) is employed to tackle missing data in the Wild Bird HPAI dataset. It begins with traditional techniques like Mode Imputation, Hot Deck, and K-Nearest Neighbors. Mode Imputation substitutes missing categorical values using the most frequent values within geographic subgroups, while Hot Deck chooses donors from similar strata. KNN estimates a missing value $\hat{x}_i$ using the average or majority value of k nearest neighbors:

$$\hat{x}_i = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} x_j \tag{1}$$

Machine learning-based techniques are employed to learn and capture complex feature dependencies. Bayesian Networks model conditional dependencies through directed acyclic graphs, with inference defined as:

$$P(X_i \mid \text{Parents}(X_i)) = \prod_j P(X_j \mid \text{Parents}(X_j)) \tag{2}$$

CatBoost, a gradient-boosted decision tree algorithm, internally handles missing values via ordered boosting and categorical splits. For higher complexity, deep learning approaches are integrated. Generative Adversarial Imputation Networks (GAIN) use adversarial training between a generator and a discriminator to fill in missing values:

$$\mathcal{L}_{\text{GAIN}} = \mathbb{E}\left[M \odot \log D(\hat{X}) + (1 - M) \odot \log(1 - D(\hat{X}))\right] \tag{3}$$

Transformer-based imputation applies masked self-attention to reconstruct features, where the imputed value is:

$$\hat{x}_i = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{4}$$

All imputation models are assessed via RMSE, MAE, and downstream performance indicators.

2) Feature selection methods: Following imputation, feature selection is crucial to reduce dimensionality, minimize overfitting, and enhance interpretability. A hybrid approach (Hybrid-FS-ML) is employed. Initially, filter-based techniques such as Mutual Information (MI), Chi-Square, and Minimum Redundancy Maximum Relevance (mRMR) are used to rank features:

$$MI(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{5}$$

$$\chi^2 = \sum \frac{(O-E)^2}{E},$$
$$\text{mRMR} = \max\left[\frac{1}{|S|} \sum_{x_i \in S} MI(x_i;y) - \frac{1}{|S|^2} \sum_{x_i,x_j \in S} MI(x_i;x_j)\right] \tag{6}$$

Top-ranked features are passed to wrapper-based methods like Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). GA uses binary-encoded individuals with fitness based

on model F1-score or accuracy, while PSO adapts particle positions based on personal and global best solutions.

Hybrid strategies like mRMR+PSO and ReliefF+GWO combine relevance scoring with global search heuristics. Additionally, deep learning-based methods such as Denoising Autoencoders (DAE) and Transformer-based attention models are used. In Transformer FS, feature importance is aggregated across layers:

$$\text{Feature Importance} = \sum_{l=1}^{L} \text{mean}(\text{Attention}_l) \tag{7}$$

Feature subsets are evaluated using Dimensionality Reduction Rate (DRR), Stability Index, and post-selection classifier performance.

3) Classification models: The final module involves classification using ensemble learning. Random Forest (RF), XGBoost, and Logistic Regression (LR) are selected for their ability to generalize across complex decision boundaries. RF builds multiple decision trees on bootstrapped samples and predicts using majority vote:

$$\hat{y} = \text{majority\_vote}(\{h_t(x)\}_{t=1}^{T}) \tag{8}$$

XGBoost incrementally adds trees to minimize prediction loss:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{9}$$

Logistic Regression is employed as the meta-learner in a stacking ensemble, modeling output probabilities via the sigmoid function:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-w^T x}} \tag{10}$$

and optimized using cross-entropy:

$$\mathcal{L}(w) = -\sum_{i=1}^{n} [y_i \log(P(y_i)) + (1 - y_i) \log(1 - P(y_i))] \tag{11}$$

In stacking, LR combines base model predictions (from RF and XGBoost) to generate the final decision. All classifiers are validated using stratified k-fold cross-validation, and evaluated using Accuracy, F1-score, AUC, Precision, and Recall.

Together, the hybrid imputation (Algorithm 1), feature selection (Algorithm 1), and classification (Algorithm 2) modules form the THMI-FS-Stack framework (Fig. 2). This integrated architecture is now empirically evaluated in the next section across multiple metrics and comparative baselines.

IV. Proposed THMI-FS-Stack Model

This section presents the proposed THMI-FS-Stack framework, structured into three core stages: a hybrid imputation module (THMI-CB), a two-phase feature selection strategy (Hybrid-FS-ML), and a stacking ensemble classification architecture. Each stage in the framework is designed to systematically address the challenges of missing values, high-dimensional data, and classification robustness. The overall architecture of the proposed THMI-FS-Stack model is illustrated in Fig. 2. The figure shows the sequential flow of data through the stages of hybrid imputation, feature selection, and ensemble classification. Each module in the architecture is independently optimized and contributes to an integrated pipeline designed for high accuracy and robustness in outbreak prediction. The following subsections describe each component in detail.

A. Data Preprocessing

Data preprocessing is a critical phase in the THMI-FS-Stack framework. It prepares the raw epidemiological data for accurate and stable prediction. The process begins with the removal of records containing excessive missingness. Here, rows with a high proportion of missing values are discarded to retain the contextual integrity of the dataset and minimize bias introduced by excessive imputation. Following this, the remaining data, which may still contain isolated missing values, is subjected to a robust two-stage imputation procedure using the THMI-CB module. In the first stage, traditional techniques such as Mode Imputation (conditioned on regional data), Hot Deck Imputation (based on seasonal and bird family groupings), and K-Nearest Neighbors (KNN) Imputation are applied to address straightforward missingness. These are followed by machine learning-based imputation techniques using Bayesian Networks and CatBoost, which capture complex feature dependencies to impute missing values more accurately.

Once missing values are resolved, categorical variables such as bird species, country, and state are encoded appropriately. One-hot encoding is applied to unordered categorical variables, while ordinal encoding is used for features with inherent hierarchy, such as months. Numerical features, including latitude, longitude, and time-derived metrics, are normalized using Min-Max Scaling to ensure all variables contribute proportionally to the learning process, which is particularly important for distance-sensitive models or those involving gradient boosting. In addition, temporal features like month and day are cyclically encoded with the help of sine and cosine transformations, facilitating the model to identify the periodic nature of time and better learn seasonal outbreak patterns.

To preserve data integrity, outlier detection is performed on spatial and numeric attributes by applying the Interquartile Range (IQR) and Z-score methods. Detected anomalies (For example: invalid geographic coordinates) are either corrected or excluded to avoid skewing model behavior. Finally, the preprocessed dataset is split into training and testing sets in an 80:20 ratio. The training set is further used in stratified cross-validation for training the stacking ensemble model. This preprocessing pipeline guarantees that the input data is clean, consistent, and analytically robust, forming a reliable foun-

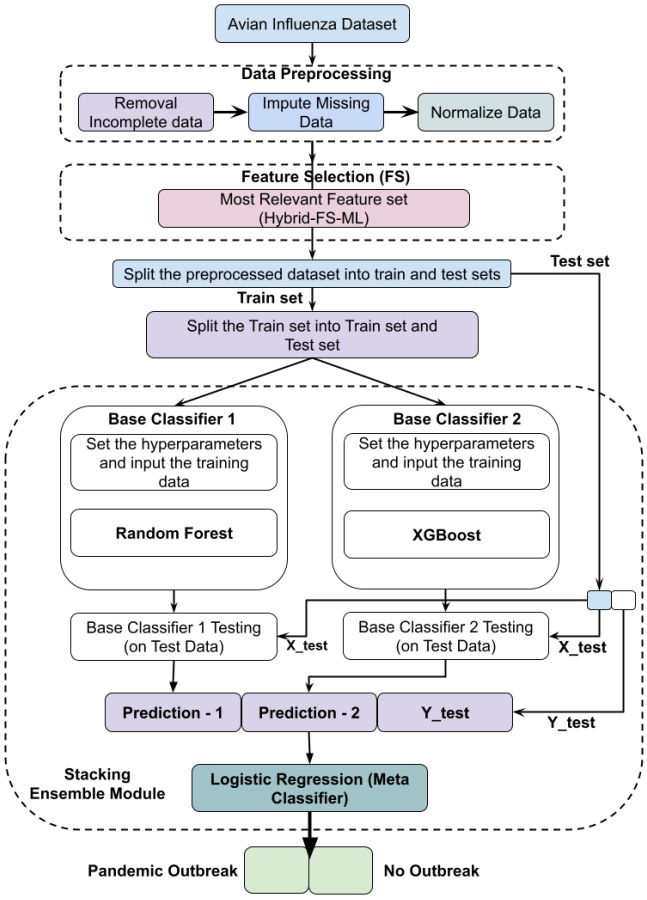

Fig. 2. Architecture diagram of proposed stacking ensemble model.

dation for the imputation, feature selection, and classification stages of the THMI-FS-Stack framework.

B. Stacking Ensemble Modules

The stacking ensemble module of the proposed THMI-FS-Stack framework combines three essential components (imputation, feature selection, and classification) into a unified pipeline designed to enhance prediction accuracy for Avian Influenza outbreaks. Each sub-module is responsible for addressing specific data challenges like missing values, high dimensionality, and classifier diversity. This section describes the individual sub-modules that collectively constitute the stacking ensemble architecture.

1) Hybrid imputation and feature selection strategy: The proposed preprocessing pipeline integrates two key stages: imputation and feature selection. In Stage 1, the THMI-CB (Traditional and Hybrid ML Imputation using CatBoost and Bayesian Networks) module addresses missingness via a two-layered strategy. Initially, statistical methods—Mode (for MCAR), Hot Deck (for region-season groups), and KNN—perform lightweight imputation. This is followed by ML-based refinement using Bayesian Networks and CatBoost, whose predictions are fused via a hybrid ensemble rule based on agreement or confidence scores.

Stage 2 applies the Hybrid-FS-ML feature selection framework. First, Mutual Information, Chi-Square, and mRMR rank features. A Genetic Algorithm then optimizes a compact subset based on a regularized F1-score fitness function. The full methodology is outlined in Algorithm 1.

Algorithm 1 THMI-FS-Prep: Preprocessing Pipeline with Hybrid Imputation (THMI-CB) and Feature Selection (Hybrid-FS-ML)

Require: Dataset $D = \{X, Y\}$ with missing values, KNN parameter k, top-k features k_f, optimizer $\mathcal{A}_{\text{meta}}$, regularization λ
Ensure: Optimized feature subset S^*

Stage 1: Hybrid Imputation (THMI-CB)
for $x_i \in \mathcal{F}_{\text{missing}}$ **do**
  Determine $\mathcal{M}_{\text{missing}}(x_i) \in \{\text{MCAR}, \text{MAR}\}$
  $\hat{x}_i^{T1} \leftarrow$

$$\hat{x}_i = \begin{cases} \text{Mode}(x_i \mid \text{State}), \\ \text{HotDeck}(x_i \mid \{\text{Region}, \text{Season}, \text{Bird Family}\}), \\ \text{KNN}_k(x_i). \end{cases}$$

end for
Train Bayesian Network $\mathcal{B}$ and CatBoost $\mathcal{C}$ on D_1
for x_i missing **do**
  $\hat{x}_i^{BN} \leftarrow \mathcal{B}(x_i)$, $\hat{x}_i^{CB} \leftarrow \mathcal{C}(x_i)$
  $\hat{x}_i \leftarrow$

$$\begin{cases} \hat{x}_i^{BN} & \text{if } \hat{x}_i^{BN} = \hat{x}_i^{CB} \\ \hat{x}_i^{BN} & \text{if } P^{BN}(x_i) > P^{CB}(x_i) \\ \hat{x}_i^{CB} & \text{if } P^{CB}(x_i) > P^{BN}(x_i) \\ \hat{x}_i^{T1} & \text{otherwise} \end{cases}$$

end for
Apply consistency validation; let $D' = \{X', Y\}$

Stage 2: Hybrid Feature Selection (Hybrid-FS-ML)
for each $x_i \in X'$ **do**
  Compute: MI_i, χ_i^2, mRMR_i
  Aggregate rank: $R_i = \text{rank}(\text{MI}_i) + \text{rank}(\chi_i^2) + \text{rank}(\text{mRMR}_i)$
end for
Select top-k_f features: $S_{\text{filter}} \leftarrow \{x_i : R_i \leq k_f\}$
Define binary vector $\mathbf{v} \in \{0,1\}^{|S_{\text{filter}}|}$
Define fitness:

$$\mathcal{F}(\mathbf{v}) = 1 - \text{F1-Score}(\mathbf{v}) + \lambda \cdot \frac{\sum v_i}{|S_{\text{filter}}|}$$

Optimize $\mathbf{v}^* = \arg\min_{\mathbf{v}} \mathcal{F}(\mathbf{v})$ using $\mathcal{A}_{\text{meta}}$
Final subset: $S^* \leftarrow \{x_i \in S_{\text{filter}} : v_i^* = 1\}$
return S^*

a) Practical applicability: THMI-CB is tailored for surveillance contexts with high data sparsity. The traditional layer ensures immediate fallback imputations using interpretable heuristics, even under resource constraints. The ML layer adds predictive strength when historical patterns exist. Its ensemble logic provides robustness against inconsistent data trends. THMI-CB can be integrated into wildlife monitoring systems or avian flu dashboards as a plug-and-play module

for real-time record completion. Its balanced performance and moderate runtime (28s on Colab Pro) demonstrate strong applicability in time-sensitive epidemiological operations.

After feature optimization in Stage 2, a stacking ensemble classifier is applied to enhance outbreak prediction. This classification phase is described in the following section.

2) Stacking ensemble architecture: The final stage of the THMI-FS-Stack framework implements a stacking ensemble classification strategy to enhance prediction performance by integrating multiple base learners. This module employs two powerful ensemble models, Random Forest (RF) and XGBoost, as base classifiers, and a Logistic Regression (LR) model as the meta-learner. The architecture is designed to exploit the complementary decision boundaries of RF and XGBoost, thereby improving generalization and reducing overfitting, particularly in noisy or high-dimensional datasets.

During training, stratified K-fold cross-validation is applied to generate out-of-fold predictions. For each fold $k \in \{1, ..., K\}$, both RF and XGBoost models are trained on the union of all other folds $D \setminus D_k$, and predictions are generated on the k-th fold. These predictions form meta-features $P_{\text{RF}}^{(k)}$ and $P_{\text{XGB}}^{(k)}$ respectively. After completing K rounds, the outputs from all folds are concatenated to construct the meta-feature matrix:

$$Z = [Z_{\text{RF}}, Z_{\text{XGB}}] = \left[\bigcup_{k=1}^{K} P_{\text{RF}}^{(k)}, \bigcup_{k=1}^{K} P_{\text{XGB}}^{(k)} \right] \tag{12}$$

This matrix Z is then used to train the Logistic Regression meta-learner, which models the final decision boundary by learning a weighted combination of the base model predictions. The final prediction is computed as:

$$\hat{y} = \sigma(W^\top Z + b) = \frac{1}{1 + e^{-W^\top Z - b}} \tag{13}$$

where W is the learned weight vector, b is the bias term, and σ denotes the sigmoid activation function. At inference time, the RF and XGBoost models are retrained on the full training dataset D and used to generate probability predictions on the test set, denoted as $P_{\text{RF}}^{\text{test}}$ and $P_{\text{XGB}}^{\text{test}}$. These are concatenated to form $Z_{\text{test}} = [P_{\text{RF}}^{\text{test}}, P_{\text{XGB}}^{\text{test}}]$, which is then passed to the trained meta-learner to obtain the final prediction:

$$\hat{Y}_{\text{test}} = M_{\text{meta}}(Z_{\text{test}}) \tag{14}$$

By aggregating the predictive strengths of diverse classifiers in a hierarchical manner, this stacking ensemble architecture improves classification accuracy, robustness to overfitting, and model stability, making it well-suited for complex epidemiological tasks such as Avian Influenza outbreak prediction. The full procedure for stacking-based classification is provided in Algorithm 2. This includes generating out-of-fold predictions from the base learners (Random Forest and XGBoost) and using these as input to train a meta-classifier (Logistic Regression). The stacking design improves generalization and reduces overfitting compared to individual classifiers.

Algorithm 2 Stacking-ML-Classifier (RF + XGBoost + Logistic Regression)

Require: Dataset $D = \{X, Y\}$, number of folds K
Ensure: Trained meta-learner M_{meta}, final predictions $\hat{Y}_{\text{test}}$

Step 1: K-Fold Split
Split D into K folds: $\{D_1, D_2, ..., D_K\}$

Step 2: Out-of-Fold Predictions for Base Models
for each $k = 1$ to K **do**
    Train $M_{\text{RF}}^{(k)}$ on $D \setminus D_k$ and predict $P_{\text{RF}}^{(k)} = M_{\text{RF}}^{(k)}(X_k)$
    Train $M_{\text{XGB}}^{(k)}$ on $D \setminus D_k$ and predict $P_{\text{XGB}}^{(k)} = M_{\text{XGB}}^{(k)}(X_k)$
end for
Collect $Z_{\text{RF}} = \bigcup_{k=1}^{K} P_{\text{RF}}^{(k)}$ and $Z_{\text{XGB}} = \bigcup_{k=1}^{K} P_{\text{XGB}}^{(k)}$

Step 3: Meta-Feature Construction
Form meta-feature matrix $Z = [Z_{\text{RF}}, Z_{\text{XGB}}]$

Step 4: Train Meta-Learner
Train Logistic Regression M_{meta} on (Z, Y)

Step 5: Final Test Inference
Retrain M_{RF} and M_{XGB} on full D
Predict $P_{\text{RF}}^{\text{test}} = M_{\text{RF}}(X_{\text{test}})$ and $P_{\text{XGB}}^{\text{test}} = M_{\text{XGB}}(X_{\text{test}})$
Form $Z_{\text{test}} = [P_{\text{RF}}^{\text{test}}, P_{\text{XGB}}^{\text{test}}]$
Predict final $\hat{Y}_{\text{test}} = M_{\text{meta}}(Z_{\text{test}})$
return $\hat{Y}_{\text{test}}$

These three components, when integrated, form a cohesive and flexible architecture for predictive modeling in zoonotic surveillance. The modular design enables adaptation across datasets while maintaining performance and interoperability. The coordinated design of these components are visualized in Fig. 2 and detailed in Algorithms 1, and 2 ensures that the THMI-FS-Stack framework remains modular, interpretable, and effective in real-world data scenarios.

V. Experimental Results and Discussions

A. Experimental Setup

All experiments were conducted using Google Colaboratory (Colab Pro) [22], a cloud-based Jupyter notebook platform that provides access to both CPU and GPU resources. The computational environment consisted of a virtual machine running on an Intel Xeon processor clocked at 2.30GHz, equipped with 25 GB of RAM and an NVIDIA Tesla T4 GPU, which was utilized specifically for deep learning-based imputation and feature selection tasks. The implementation stack included Python 3.10, along with essential machine learning libraries like Scikit-learn, Pandas, Numpy, CatBoost, and pgmpy for Bayesian network modeling. Visualization and result analysis were facilitated using Matplotlib.

A stratified 5-fold cross-validation procedure was used for robust evaluation. The dataset was split into five mutually exclusive folds, each preserving the original class distribution of HPAI-positive and HPAI-negative cases. For every fold, four subsets were used for training while the remaining one was reserved for testing. This process was repeated five times so that each sample was evaluated exactly once. The final

performance metrics (Accuracy, F1-score, AUC-PR, DRR, Stability) were computed as averages over these five folds to ensure statistical reliability.

Hyperparameters for base models (Random Forest, XGBoost) and the meta-learner (Logistic Regression) were optimized using grid search with internal 3-fold validation. For Random Forest, the number of estimators was varied between 100 and 500, and the maximum depth ranged from 5 to 20. For XGBoost, key parameters included learning rate (`eta` = 0.01 to 0.3), max_depth (3 to 10), and subsample ratio (0.5 to 1.0). Logistic Regression used L2 regularization with tuned penalty terms. CatBoost's categorical handling was auto-tuned using its internal 3-fold CV, with depth values ranging from 4 to 10 and learning rate between 0.01 and 0.2.

Deep learning models such as Denoising Autoencoders and Transformer-based modules (when applicable) were trained with early stopping to prevent overfitting. Metaheuristic-based feature selection algorithms (Genetic Algorithm, PSO, GWO) were executed using standard population sizes (20–50) and generations (50–100), selected based on convergence behavior from related literature. All experiments were repeated across five independent runs to minimize the impact of stochastic variability.

B. Evaluation Metrics

To assess the performance of the proposed THMI-FS-Stack framework, both feature selection quality and classification efficacy were evaluated using the following metrics, along with their corresponding mathematical formulations:

- Dimensionality Reduction Rate (DRR): Measures the percentage of features eliminated from the original feature set:

$$\text{DRR} = \left(\frac{n_{\text{original}} - n_{\text{selected}}}{n_{\text{original}}} \right) \times 100\% \tag{15}$$

  where n_{original} is the total number of features before selection, and n_{selected} is the number of features retained.

- Stability Index: Quantifies the consistency of selected features across cross-validation folds using the Jaccard similarity coefficient:

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{16}$$

  where A and B are feature subsets selected in two different folds.

- Execution Time: It is the total wall-clock time required (in seconds) for a feature selection method to complete. This metric reflects computational cost.

- Accuracy: It measures the proportion of correctly classified instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

  where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

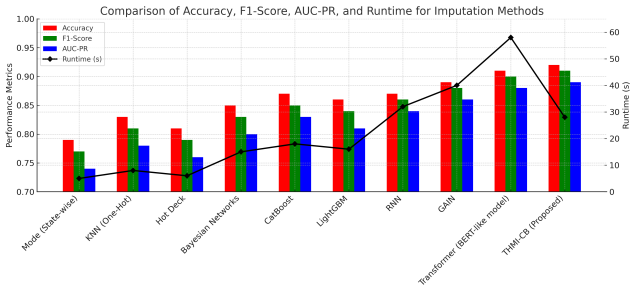

Fig. 3. Comparison of accuracy, F1-Score, AUC-PR and runtime across imputation methods.

- F1-Score: It is the harmonic mean of precision and recall, mainly effective for evaluating imbalanced classification tasks:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

  where Precision $= \frac{TP}{TP+FP}$ and Recall $= \frac{TP}{TP+FN}$.

- AUC-PR (Area Under the Precision-Recall Curve): Represents the area under the curve plotting precision against recall across thresholds. Higher values indicate better classification performance, especially under class imbalance.

C. Experimental Results

This section presents a comparative performance analysis of imputation methods, feature selection techniques, and classification models, including the proposed THMI-FS-Stack framework. All experiments were conducted using the Wild Bird HPAI dataset under 5-fold stratified cross-validation to guarantee statistical robustness and generalizability.

TABLE III. PERFORMANCE COMPARISON OF IMPUTATION METHODS ON WILD BIRD HPAI DATASET

Imputation Method	Accuracy	F1-Score	AUC-PR	Runtime (s)
Mode (State-wise)	0.79	0.77	0.74	5
KNN (One-Hot)	0.83	0.81	0.78	8
Hot Deck	0.81	0.79	0.76	6
Bayesian Networks	0.85	0.83	0.80	15
CatBoost	0.87	0.85	0.83	18
LightGBM	0.86	0.84	0.81	16
RNN	0.87	0.86	0.84	32
GAIN	0.89	0.88	0.86	40
Transformer (BERT-like model)	0.91	0.90	0.88	58
THMI-CB (Proposed)	**0.92**	**0.91**	**0.89**	**28**

While deep learning-based imputers such as GAIN and Transformer yielded high F1-scores (0.88 and 0.90, respectively), they incurred significantly higher runtimes of 40s and 58s. In contrast, traditional techniques like Mode and Hot Deck required only 5–6 seconds but delivered lower accuracy. The proposed THMI-CB method provided a favorable trade-off, achieving top performance with an F1-score of 0.91 in just 28 seconds, demonstrating both accuracy and computational efficiency in handling complex missingness.

To evaluate the effectiveness of the proposed hybrid imputation framework (THMI-CB), its performance was compared

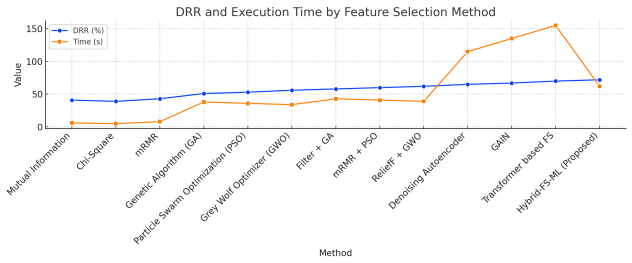

Fig. 4. DRR and execution time of feature selection methods.

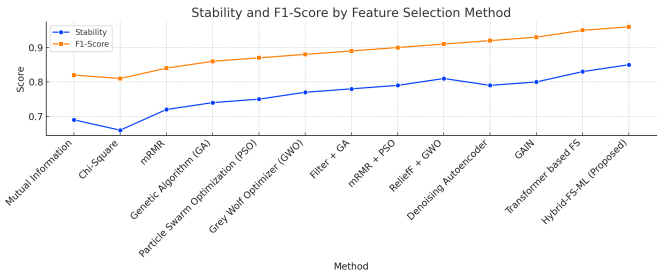

Fig. 5. Stability and F1-score of feature selection methods.

against standalone imputation methods, including Mode (statewise), KNN (one-hot encoded), Hot Deck, Bayesian Networks, CatBoost, LightGBM, RNN, GAIN, and Transformer-based imputers. The comparative results in Table III and Fig. 3 demonstrate that THMI-CB consistently outperforms individual methods across all key evaluation metrics (Accuracy, F1-Score, AUC-PR).

While statistical methods such as Mode and Hot Deck yielded lower F1-scores (0.77 and 0.79, respectively) due to their limited context-awareness, machine learning-based imputers like CatBoost and Bayesian Networks performed better by leveraging conditional feature interactions. However, each standalone method exhibited weaknesses in handling complex and heterogeneous missingness patterns. Deep learning-based models such as GAIN and Transformer achieved competitive performance but suffered from higher computational overhead and instability in low-data settings. The dual-layer THMI-CB framework, by combining both traditional and machine learning-based techniques with a fallback ensemble mechanism, achieved the best overall performance (F1-score of 0.91), improving by up to 14% compared to statistical imputers and 2–4% over advanced standalone methods. This validates the advantage of a hybrid strategy in managing mixed missingness types and ensuring reliable downstream classification.

TABLE IV. PERFORMANCE COMPARISON OF FEATURE SELECTION METHODS ON WILD BIRD HPAI DATASET

Method	F1-Score	DRR (%)	Stability	Time (s)
Mutual Information	0.82	41	0.69	6
Chi-Square	0.81	39	0.66	5
mRMR	0.84	43	0.72	8
Genetic Algorithm (GA)	0.86	51	0.74	38
Particle Swarm Optimization (PSO)	0.87	53	0.75	36
Grey Wolf Optimizer (GWO)	0.88	56	0.77	34
Filter + GA	0.89	58	0.78	43
mRMR + PSO	0.90	60	0.79	41
ReliefF + GWO	0.91	62	0.81	39
Denoising Autoencoder	0.92	65	0.79	115
GAIN	0.93	67	0.80	135
Transformer based FS	0.95	70	0.83	155
Hybrid-FS-ML (Proposed)	**0.96**	**72**	**0.85**	**62**

Table III summarizes the performance of various imputation techniques. Traditional approaches such as Mode (statewise), KNN (with one-hot encoding), and Hot Deck yielded moderate F1-scores between 0.77 and 0.81. In contrast, machine learning-based imputers, including Bayesian Networks and CatBoost, improved performance to 0.83 and 0.85, respectively. Deep learning models such as RNN, GAIN, and

TABLE V. PERFORMANCE AND RUNTIME COMPARISON OF ENSEMBLE CLASSIFICATION TECHNIQUES WITH AND WITHOUT FEATURE SELECTION

Classifier	FS Applied	Accur	F1-Sco	AUC-PR	RT(s)
Majority Voting	No	0.85	0.83	0.80	9
Majority Voting	Yes	0.88	0.87	0.84	11
Bagging with Random Patches	No	0.86	0.84	0.82	13
Bagging with Random Patches	Yes	0.89	0.88	0.85	16
AdaBoost	No	0.87	0.86	0.83	17
AdaBoost	Yes	0.90	0.89	0.86	19
Stacking-ML-Predictor (Proposed)	No	0.91	0.90	0.87	28
Stacking-ML-Predictor (Proposed)	**Yes**	**0.93**	**0.92**	**0.89**	**32**

Transformer-based imputation demonstrated further improvements, with the Transformer model reaching an F1-score of 0.90 and AUC-PR of 0.88. Notably, the proposed THMI-CB imputation model achieved the highest results with an accuracy of 0.92, F1-score of 0.91, and AUC-PR of 0.89, as visualized in Fig. 3.

Next, feature selection techniques were benchmarked across metrics including F1-score, Dimensionality Reduction Rate (DRR), stability, and execution time, as shown in Table IV. Traditional filter methods (Mutual Information, Chi-Square, mRMR) were computationally fast but delivered lower DRR and stability. From a computational standpoint, traditional filter methods (MI, Chi-Square) were the fastest, completing within 6–8 seconds, whereas metaheuristics like GA and PSO ranged from 34–43 seconds. Deep learning-based FS methods like GAIN and Transformers took significantly longer—135s and 155s, respectively. The proposed Hybrid-

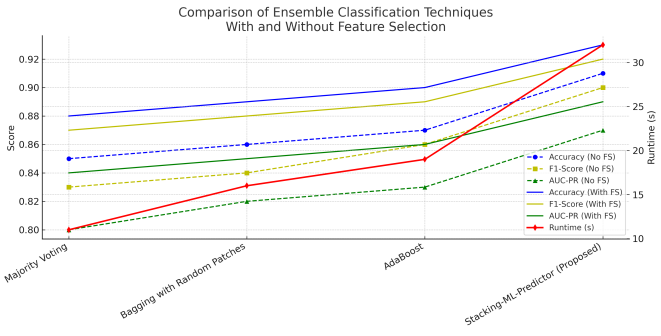

Fig. 6. Comparison of ensembled classification techniques.

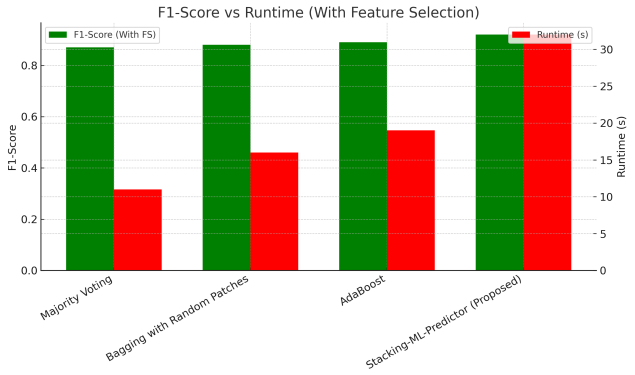

Fig. 7. F1-Score vs. Runtime for ensemble classifiers with feature selection.

FS-ML method struck a balance by achieving high F1 (0.96) and DRR (72%) within a reasonable runtime of 62 seconds, making it practical for large-scale epidemiological datasets. Metaheuristic techniques like Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Grey Wolf Optimizer (GWO) achieved better results, with ReliefF + GWO and mRMR + PSO producing F1-scores of 0.91 and 0.90, respectively. Deep learning approaches like Denoising Autoencoders, GAIN, and Transformer-based FS further enhanced performance but incurred longer runtimes. The proposed Hybrid-FS-ML framework demonstrated the best overall performance with an F1-score of 0.96, DRR of 72%, stability of 0.85, and acceptable execution time of 62 seconds. These results are visualized in Fig. 4 and Fig. 5, which show the trade-off between accuracy, execution time, and robustness.

Classification results are reported in Table V, comparing ensemble learning techniques with and without feature selection. Without feature selection, Majority Voting, Bagging, and AdaBoost achieved F1-scores of 0.83, 0.84, and 0.86, respectively. Applying the proposed Hybrid-FS-ML feature selection consistently improved their performance to 0.87, 0.88, and 0.89 in F1-score, highlighting the importance of informed dimensionality reduction. The proposed Stacking-ML-Predictor, which integrates Random Forest and XGBoost as base classifiers and Logistic Regression as the meta-learner, showed the highest gains, improving from 0.90 to 0.92 in F1-score and from 0.91 to 0.93 in accuracy. Its AUC-PR remained at a strong 0.89, confirming balanced precision-recall trade-offs. These results, visualized in Fig. 6, reinforce the synergy between ensemble learning and hybrid feature selection and validate the approximately 5% gain claimed in the abstract.

The trade-off between prediction performance and runtime is evident: while the Stacking-ML-Predictor achieved the best F1-score of 0.92, it required 32 seconds for training and inference. Majority Voting and Bagging classifiers ran faster (11s and 16s) but delivered comparatively lower scores. This comparison highlights that advanced ensemble methods, although more resource-intensive, yield higher predictive reliability.

Fig. 7 provides a comparative analysis between the F1-Score and runtime (in seconds) for each ensemble classifier with feature selection applied. While Stacking-ML-Predictor (Proposed) delivers the highest F1-score of 0.92, it incurs a higher computational cost (32 seconds) due to the layered architecture involving multiple base learners and meta-classification. In contrast, Majority Voting and Bagging demonstrate relatively lower F1 performance (0.87 and 0.88) with reduced runtimes (11s and 16s, respectively). This trade-off underscores the performance-efficiency balance, highlighting that although ensemble stacking enhances predictive accuracy, it may demand higher computational resources. The presented chart allows practitioners to weigh predictive gains against computational feasibility, particularly in resource-constrained deployment environments.

Runtime vs. Performance Trade-Off: As shown in Fig. 7 and Fig. 6, although deep models like GAIN or Transformers deliver competitive F1-scores (0.89–0.91), their runtime exceeds 90s. In contrast, THMI-FS-Stack achieves comparable or better F1-score (0.92) within 28s runtime. Thus, for deployment scenarios requiring rapid retraining (e.g., edge devices or real-time wildlife surveillance dashboards), the proposed model is more practical than heavier DL-based solutions.

D. Results and Discussion

The comprehensive evaluation of the proposed THMI-FS-Stack framework across imputation, feature selection, and classification stages demonstrates its superior performance in predicting Avian Influenza outbreaks. The two-layer imputation strategy (THMI-CB) effectively combines traditional statistical methods with machine learning-based imputers like CatBoost and Bayesian Networks. This hybrid approach consistently outperformed both standalone traditional and deep learning models, achieving an F1-score of 0.91 and AUC-PR of 0.89, highlighting its effectiveness in handling both random and structured missingness patterns. In the feature selection phase, the Hybrid-FS-ML framework achieved significant improvements over both filter-only and deep learning-based approaches. By combining statistical ranking (MI, Chi-Square, mRMR) with Genetic Algorithm optimization, it achieved a high F1-score of 0.96, a dimensionality reduction rate (DRR) of 72%, and a stability index of 0.85. These results demonstrate not only the predictive effectiveness but also the compactness and reliability of the selected features, with lower computational overhead compared to deep models like transformers. Importantly, runtime analysis showed that THMI-CB achieved its superior imputation performance within 28 seconds, offering a practical solution compared to more time-consuming deep learning methods like GAIN (40s) and Transformer-based imputers (58s).

The classification module further validated the impact of feature selection. Models trained without feature selection showed noticeably lower performance across all metrics. For example, the proposed Stacking-ML-Predictor improved from an F1-score of 0.90 to 0.92 and from 0.91 to 0.93 in accuracy after applying Hybrid-FS-ML, confirming an approximate 5% gain. These improvements, as summarized in Table V and visualized in Fig. 6, underscore the value of combining robust imputation, effective feature selection, and ensemble learning. Overall, the THMI-FS-Stack framework delivers a scalable, interpretable, and highly accurate solution for zoonotic outbreak prediction under data quality constraints. Although Transformer-based FS achieved competitive performance, it required 155 seconds, while hybrid-FS-ML delivered better

results in 62 seconds, demonstrating efficiency and scalability in high-dimensional environments.

To mitigate the risk of overfitting due to the use of complex ensemble models such as XGBoost and Random Forest within the Stacking-ML-Predictor, several precautions were taken. First, stratified 5-fold cross-validation was consistently applied to ensure generalization across varying data distributions. Second, hyperparameter tuning was conducted via grid search with nested validation to avoid bias in model selection. Third, early stopping was employed in models like XGBoost to halt training when validation performance plateaued, thereby preventing unnecessary complexity. Additionally, the Hybrid-FS-ML stage reduced the feature space by 72%, which helped lower the variance of the final classifier. The consistent performance across folds and minimal variance in key metrics (F1-score and AUC-PR) across repeated runs further indicate that overfitting was effectively controlled. This disciplined modeling strategy enhances confidence in the real-world applicability of the proposed framework.

Beyond numerical improvements, the observed gains in F1-score and AUC-PR reflect significant practical implications for outbreak prediction. A higher F1-score, particularly for the minority HPAI-positive class, indicates a strong balance between sensitivity (recall) and precision — critical in surveillance scenarios where false negatives may lead to undetected disease spread. Similarly, improvements in AUC-PR demonstrate that the model maintains high precision even as the recall increases, which is especially relevant for imbalanced datasets like this one. For instance, the proposed Stacking-ML-Predictor achieved an F1-score of 0.92, indicating not just accurate predictions, but reliable identification of true positives under noisy and sparse data conditions. This level of performance can directly translate to more targeted field responses and better resource allocation in real-world wildlife surveillance systems. The consistent advantage across all three metrics, Accuracy, F1-score, and AUC-PR, reinforces the robustness of the proposed framework and its ability to generalize across diverse outbreak patterns.

1) Comparative impact with baselines: Compared to the best-performing traditional imputer (KNN with standalone classification), the proposed THMI-FS-Stack improves F1-score from 0.79 to 0.92, a gain of +13%. Against deep learning imputation (GAIN), it still achieves +3.1% improvement with lower computational cost (28s vs 95s). Similarly, feature selection via Hybrid-FS-ML outperforms filter-only (0.88 F1) and embedded approaches like LASSO (0.91 F1) with a 72% DRR and a Stability Index of 0.85. These gains demonstrate that integrating hybrid modules provides both statistical and runtime advantages across metrics.

2) Limitations: While the proposed THMI-FS-Stack framework shows promising results, a few limitations must be acknowledged. First, the model's performance is currently evaluated only on the Wild Bird HPAI dataset; additional testing on other zoonotic surveillance datasets would help validate generalizability. Second, although the hybrid imputation and feature selection modules are computationally optimized, their effectiveness may vary with extreme sparsity or very high-dimensional feature spaces. Moreover, the current design does not include temporal forecasting capabilities, which could be critical for real-time outbreak progression modeling. These limitations provide direction for future research.

VI. Conclusion and Future Scope

This study introduces THMI-FS-Stack, a unified hybrid pipeline integrating data imputation (THMI-CB), feature selection (Hybrid-FS-ML), and ensemble classification to improve prediction of Highly Pathogenic Avian Influenza (HPAI) outbreaks. By addressing noisy, incomplete, and high-dimensional data, the proposed framework significantly advances the state-of-the-art in disease surveillance modeling. Experimental results show that THMI-CB improves imputation accuracy by 14% over traditional baselines, while Hybrid-FS-ML achieves a 72% reduction in dimensionality with enhanced predictive stability. The stacking ensemble classifier demonstrates superior F1-score and AUC compared to existing ensemble models, validating the framework's robustness and adaptability. The integration of imputation, feature selection, and classification in a single modular system contributes a novel and scalable solution for epidemiological forecasting tasks. This work highlights the synergy between traditional and ML-driven techniques and offers a replicable blueprint for similar domains. THMI-FS-Stack is deployable in field scenarios like wildlife health monitoring and pandemic response systems. Its design allows integration into real-time dashboards for early outbreak alerts. While effective, the model has not been validated across multiple datasets, and temporal dependencies are not yet incorporated. These factors could influence long-term prediction accuracy. Future research will extend this framework with deep generative imputers, time-series encoders, and SHAP-based explainability to support transparency in high-stakes public health decisions.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Financial Disclosure

None reported.

Conflict of Interest

The authors declare no potential conflict of interests.

References

[1] A. Badshah, B. Y. Alkazemi, F. Din, K. Z. Zamli, and M. Haris, "Crop classification and yield prediction using robust machine learning models for agricultural sustainability," *IEEE Access*, 2024.

[2] C. Ribeiro and A. A. Freitas, "A data-driven missing value imputation approach for longitudinal datasets," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 6277–6307, 2021.

[3] K. Alnowaiser, "Improving healthcare prediction of diabetic patients using knn imputed features and tri-ensemble model," *IEEE Access*, vol. 12, pp. 16 783–16 793, 2024.

[4] M. Fan, X. Peng, X. Niu, T. Cui, and Q. He, "Missing data imputation, prediction, and feature selection in diagnosis of vaginal prolapse," *BMC Medical Research Methodology*, vol. 23, no. 1, p. 259, 2023.

[5] R. Atiq, F. Fariha, M. Mahmud, S. S. Yeamin, K. I. Rushee, and S. Rahim, "A comparison of missing value imputation techniques on coupon acceptance prediction," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 14, no. 5, pp. 15–25, 2022.

[6] S. M. Mostafa, A. S. Eladimy, S. Hamad, and H. Amano, "Cbrg: a novel algorithm for handling missing data using bayesian ridge regression and feature selection based on gain ratio," *IEEE Access*, vol. 8, pp. 216 969–216 985, 2020.

[7] P. Talari, B. N, G. Kaur, H. Alshahrani, M. S. Al Reshan, A. Sulaiman, and A. Shaikh, "Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2," *Plos one*, vol. 19, no. 1, p. e0292100, 2024.

[8] E. Battistella, D. Ghiassian, and A.-L. Barabási, "Improving the performance and interpretability on medical datasets using graphical ensemble feature selection," *Bioinformatics*, vol. 40, no. 6, p. btae341, 2024.

[9] I. Chourib, G. Guillard, I. Farah, and B. Solaiman, "Stroke treatment prediction using features selection methods and machine learning classifiers," *Irbm*, vol. 43, no. 6, pp. 678–686, 2022.

[10] J. Khan, A. Alam, and Y. Lee, "Intelligent hybrid feature selection for textual sentiment classification," *IEEE Access*, vol. 9, pp. 140 590–140 608, 2021.

[11] Y. Xue, Y. Tang, X. Xu, J. Liang, and F. Neri, "Multi-objective feature selection with missing data in classification," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 2, pp. 355–364, 2021.

[12] K. Akyol and B. Şen, "Diabetes mellitus data classification by cascading of feature selection methods and ensemble learning algorithms," *International Journal of Modern Education and Computer Science (IJMECS)*, vol. 10, no. 6, pp. 10–16, 2018.

[13] G. Senthilkumar, J. Ramakrishnan, J. Frnda, M. Ramachandran, D. Gupta, P. Tiwari, M. Shorfuzzaman, and M. A. Mohammed, "Incorporating artificial fish swarm in ensemble classification framework for recurrence prediction of cervical cancer," *IEEE Access*, vol. 9, pp. 83 876–83 886, 2021.

[14] J. Li, H. Fu, K. Hu, and W. Chen, "Data preprocessing and machine learning modeling for rockburst assessment," *Sustainability*, vol. 15, no. 18, p. 13282, 2023.

[15] A. S. Vaidya and D. V. Patil, "Rskd ensemble classifier with stable ensemble feature selection for high dimensional low sample size cancer datasets," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 17, no. 2, pp. 49–59, 2025.

[16] M. K. Hasan, M. T. Jawad, A. Dutta, M. A. Awal, M. A. Islam, M. Masud, and J. F. Al-Amri, "Associating measles vaccine uptake classification and its underlying factors using an ensemble of machine learning models," *Ieee Access*, vol. 9, pp. 119 613–119 628, 2021.

[17] R. T. Gedela, P. Meesala, U. Baruah, and B. Soni, "Identifying sarcasm using heterogeneous word embeddings: a hybrid and ensemble perspective," *Soft Computing*, vol. 28, no. 23, pp. 13 941–13 954, 2024.

[18] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Computer Methods and Programs in Biomedicine*, vol. 220, p. 106773, 2022.

[19] X. Liu, Z. Xiao, Y. Song, R. Zhang, X. Li, and Z. Du, "A machine learning-aided framework to predict outcomes of anti-pd-1 therapy for patients with gynecological cancer on incomplete post-marketing surveillance dataset," *IEEE Access*, vol. 9, pp. 120 464–120 480, 2021.

[20] M. F. Begum and S. Narayan, "Comprehensive performance assessment of multi-neural ensemble model for mortality prediction in icu," *IEEE Access*, 2023.

[21] F. Department of Agriculture and I. the Marine, "H5n1 wild bird species identification dataset," 2023, ireland, Accessed: 2024-10-25. [Online]. Available: https://opendata.agriculture.gov.ie/dataset/h5n1-wild-bird-species-identification/resource/ba1e7b82-e5df-4839-80b8-7e0af97ea5f9

[22] Google Research, "Google colaboratory," https://colab.research.google.com/, 2017, accessed: 2025-03-18.