Received: 7 August 2025 | Revised: 23 December 2025 | Accepted: 20 January 2026 | Published online: 9 February 2026

RESEARCH ARTICLE

Medinformatics
2026, Vol. 00(00) 1–12
DOI: 10.47852/bonviewMEDIN62027119

BON VIEW PUBLISHING

Advancing Clinical Diagnosis Through Comparative Analysis of Machine Learning and Transformer-Based Models for Hepatitis B Detection

Akinyemi Omololu Akinrotimi[1,*], Israel Oluwabusayo Omotosho[2], Olugbenga Olayinka Owolabi[3], Oluwaseun Adewale Olubunmi[4], Ibrahim Garba[5] and Ndie Ngalame Dionysius[6]

[1] *Department of Information Systems and Technology, Kings University, Nigeria*

[2] *Department of Management Information Systems, Bowie State University, USA*

[3] *Department of Electrical and Electronics Engineering, Adeleke University, Nigeria*

[4] *Department of Computer Engineering, Federal University Oye-Ekiti, Nigeria*

[5] *Department of Biological Sciences, Njala University, Sierra Leone*

[6] *Department of Biochemistry, University of Buea, Cameroon*

Abstract: Early and accurate diagnosis of hepatitis B remains a critical issue in low-resource healthcare settings, where reliance on straightforward laboratory tests might limit the sensitivity of diagnosis. This research explores the prediction of hepatitis B infection based on conventional biochemical markers and demographic data by four machine learning models: XGBoost, LightGBM, TabPFN, and TabKANet. The goals are twofold: first, to compare which model type is best, the Gradient Boosted Decision Trees (GBDTs) or transformer-based models, and second, to compare the feasibility of such models being included in clinical workflows. The models were evaluated against a real-world dataset of a Ghanaian hospital-based screening program. Performance was calculated in terms of accuracy, precision, recall, F1-score, and area under the curve (AUC). Results showed that transformer models, particularly TabKANet, outperformed traditional GBDT models in recall (97.0%), F1-score (65.5%), and AUC (0.92), with high ability to detect true positive cases. Confusion matrix analyses also confirmed the validity of TabKANet in minimizing false negatives, which is very much required in clinical diagnosis. These findings support the application of artificial intelligence-based technology in clinical laboratories as a second confirmatory system for disease diagnosis. The study also encourages further research into the use of transformer models in healthcare, especially in settings where confirmatory testing is unavailable.

Keywords: clinical diagnosis, gradient boosted trees, hepatitis B, machine learning, tabular transformers

1. Introduction

Hepatitis B virus (HBV) is still a significant global disease burden, particularly in low-resource settings where diagnosis is delayed or improper. In 2020, there were around 296 million HBV chronically infected persons, and there were not many who were properly diagnosed since molecular tests such as polymerase chain reaction (PCR) or imaging were not accessible [1, 2]. From a medical point of view, hepatitis B virus (HBV) is a hepatotropic DNA virus known to cause acute and chronic infection of the liver, and its mechanism of transmission is mostly due to contact with infectious blood or body fluid [3]. The virus infects hepatocytes, which leads to inflammation and subsequent damage to the hepatocytes; this can lead to cirrhosis and hepatocellular carcinoma in chronic infection. Screening and confirmation can be done through serological tests such as hepatitis B surface antigen (HBsAg), antibody to hepatitis B core antigen (anti-HB), and HBV DNA. Prevention is conducted through vaccination. Management of chronic infection is conducted through suppressive therapy with antiviral medications such as tenofovir and entecavir [4]. To provide more details, the medical tests used in diagnosing the disease are as follows:

***Corresponding author:** Akinyemi Omololu Akinrotimi, Department of Information Systems and Technology, Kings University, Nigeria. Email: akinrotimiakinyemi@ieee.org

1) Routine Serological and Liver Function Tests (LFTs): The gold standard for clinical diagnosis includes initial serologic tests (such as HBsAg, anti-HBc, antibody to hepatitis B surface antigen (anti-HBs)), in combination with liver function tests (ALT, AST, and bilirubin), for determining the extent of inflammation and subsequent damage to the liver.
2) Microfluidic Immunoassays: The chips utilize magnetic nanoparticles coated with antibodies to capture and enrich viruses in blood, combined with easy amplification in the chip for sensitive detection without relying on advanced lab equipment [5].
3) CRISPR-Cas13 L: These strips are intended for outdoor applications, enabling the visible detection of HBV DNA in less than 30 min without professional training or any equipment, making them suitable for resource-limited areas [6].
4) Photonic Crystal Biosensors Aptamer-functionalized: These are inexpensive chips allow visible colorimetric detection of HBsAg with visual observation only, without the need for sophisticated analyzers [7].
5) Paper-Based Imm Gold nanoparticles: These are used in paper dipsticks to offer colorimetric analysis for hepatitis antigens in only 15 min without using electricity or much technical knowledge [8].

Present machine learning (ML) algorithms have tremendous potential to maximize the detection of disease with readily available clinical data. Moulaei et al. [9] carried out a meta-analysis of 21 studies that used 82 ML algorithms to diagnose viral hepatitis, with 90% sensitivity and 94% specificity of support vector machines in classifying HBV. A subsequent Nigerian study employed tree-based classifiers and Support Vector Machines (SVMs) on routine laboratory results, with 90% accuracy and 0.90 area under the curve (AUC) in discriminating between HBV-positive and HBV-negative patients [10]. These reports point to the ability of ML to compete with performance by conventional immunoassays, although the majority employ conventional algorithms such as SVMs, random forests, and Gradient Boosted Decision Trees (GBDT). GBDT algorithms like XGBoost and LightGBM are today's gold standard for structured data classification due to their strength, interpretability, and scalability [11]. LightGBM, introduced by Microsoft in 2017, has since been widely validated in medical diagnostics [12], while XGBoost was the predictive modeling benchmark winner from the day it was released in 2016 [13]. While they perform effectively, these models are required to be extensively hyperparameter-tuned, and this is a task that may be time-consuming and computationally expensive in the clinical setting.

In the past few years, transformer models for tabular data have emerged, with automatic tuning and state-of-the-art or even superior performance. TabPFN, introduced as a tabular inference foundation model, utilizes a transformer model that is pre-trained on generated data and can perform Bayesian-style classification in a single forward pass [14]. A 2025 benchmark in Nature showed that TabPFN outperformed tuned GBDT models in terms of performance on datasets of up to 10,000 samples with much less runtime [15]. TabKANet, released during the latter half of 2023, is an extension of tabular transformers with Kolmogorov–Arnold networks and attention and numerical feature optimization and has demonstrated improved performance on numerous structured datasets [16]. Both are fully implemented in Python and available in pip-installable packages.

To our knowledge, there are few studies that have compared the diagnostic performance of GBDT models and cutting-edge transformer-based models—namely, TabPFN and TabKANet—for hepatitis B using routine laboratory data. This represents an important lacuna since accurate diagnosis based only on biochemical markers has the potential to empower medical laboratory scientists with artificial intelligence (AI)-driven cross-checking tools. This kind of integration could reduce diagnosis errors and accelerate clinical decision-making. The focus of this study thus goes beyond model identification and assessment to clinical relevance, which explores its application and utility in terms of its ability to reduce false negatives and its feasibility in real-world healthcare settings. This dual methodological and clinical perspective distinguishes the present work from prior ML-based hepatitis studies. This study carries out a comparative assessment of LightGBM, XGBoost, TabPFN, and TabKANet in diagnosing HBV based on traditional laboratory tests. We evaluate and train each model on performance metrics including accuracy, AUC, precision, recall, and F1-score. In addition, we examine the economic effect of deploying the model with the best performance in resource-scarce healthcare settings. By focusing on typical laboratory results and easily accessible ML software, we hope to show not only the potential but also the daily utility of AI in the diagnostic workstream. In support of this goal, we made use of an openly accessible dataset from a hepatitis B and C standard screening and treatment eligibility program in Cape Coast, Ghana [17]. This dataset includes demographic and laboratory data collected routinely in clinical practice and provides a realistic foundation for the evaluation of AI-based diagnostic models. We consider that transformer-based models, particularly TabPFN or TabKANet, will outperform traditional GBDT approaches and therefore represent an outstanding opportunity for future clinical diagnostic use. This project aims to encourage adoption by medical laboratory scientists and to guide future researchers in tool development.

2. Background and Related Work

Hepatitis B virus infection remains a major global health concern with the highest prevalence in sub-Saharan Africa and East Asia, and diagnosis typically relies on serological markers such as HBsAg and core antibody together with liver function tests, which may not identify early or asymptomatic infection without supplemental molecular testing [18]. While PCR tests are very specific, recent interpretable ML approaches have emerged as lower-cost diagnostic alternatives [19].

In the last couple of years, ML has become broadly prevalent in the healthcare field for predictive analytics, diagnosis, and treatment recommendations. Supervised ML models founded on clinical and biochemical symptoms have shown promise in disease classification, with particular reference to chronic and infectious diseases [10]. For example, logistic regression, support vector machines, and decision tree ensembles have predicted HBV status from results of routine laboratory tests with accuracies exceeding 85% in several retrospective studies [9].

Among tree-based ensemble models, GBDT have earned particular popularity in the context of structured (tabular) healthcare data. XGBoost in 2016 and LightGBM in 2017 introduced speed, regularization, and missing value handling improvements [13, 20].

There have also been several research works that successfully applied GBDT models to clinical prediction problems, such as diabetes screening, sepsis onset prediction, and hepatitis classification [21]. LightGBM also outperformed logistic regression and SVM in 2022 for hepatitis B detection from blood chemistry data with an AUC of 0.91 [16]. GBDT models typically need a lot of hyperparameter tuning, and their performance can be

sensitive to data imbalance, typical for medical data. They are poor at learning complicated interactions between features unless specifically designed into them [22]. Most recently, transformer models originally developed for natural language processing have been applied to tabular data. Transformer models, which utilize self-attention mechanisms in an effort to capture dependencies between input features, have proven highly effective in overcoming the limitations of conventional ML methods.

TabPFN (Tabular Prior-data Fitted Network), introduced in 2022, provides zero-shot classification from a pre-trained transformer pre-trained on millions of artificial tabular tasks [23]. It totally removes the need for iterative training or hyperparameter optimization and has shown strong performance on small-to-mid-sized medical datasets [15].

TabKANet, another novel model, gives Kolmogorov–Arnold networks a transformer backbone that allows it to handle continuous variables well and learn complex nonlinear relationships [21]. Relative to GBDT models, TabKANet also better performed in several benchmark tests on healthcare and finance datasets [24]. More recent discussion has also explained the appeal of TabPFN due to its zero-training framework and efficient inference time for tabular classification tasks [25].

Summary of related works on machine learning approaches to hepatitis B diagnosis is given in Table 1. Despite all these advances, direct comparisons between transformer-based models and GBDT models in the area of disease diagnosis, particularly for hepatitis B, have been fewer. Also, although ML solutions to HBV prediction have been examined, combining clinical interpretability and economic analysis is quite rare. There has also been minimal concern about how such approaches can help medical laboratory scientists in operational diagnostic procedures.

This study bridges these gaps by comparing the performance of four ML models, that is, XGBoost, LightGBM, TabPFN, and TabKANet, on a real-world hepatitis B dataset of a Ghanaian clinical screening scheme. Besides prediction performance, the goal is also to identify the potential of these models as cross-checking tools in medical labs. By establishing the most effective methodology and outlining its economic viability, the current study lends strength to the growing line of research advocating AI-based support systems for clinical diagnosis.

2.1. XGBoost model

XGBoost is a highly efficient version of the GBDT framework, created in 2016 [13]. XGBoost improves on traditional GBDT methods by using tree pruning, regularization, and parallel computation. XGBoost constructs decision trees in sequence, wherein each subsequent tree attempts to correct errors of the initial ensemble. This yields highly accurate predictive performance on structured data and prevents overfitting through L1/L2 penalties. In clinical applications, XGBoost has demonstrated strong performance in applications from risk stratification of diseases to predictions of patient outcomes [13]. The mathematical

Table 1. Summary of related works on machine learning approaches to hepatitis B diagnosis

Study	Year	Author contribution	Limitation
Chen and Guestrin [13]	2016	Introduced XGBoost, a scalable and efficient GBDT model	Requires tuning and does not adapt to unseen datasets without retraining
Ke et al. [20]	2017	Developed LightGBM, optimized for large-scale structured data	Sensitive to hyperparameter choices; does not generalize automatically
Wu et al. [19]	2022	Introduced TabPFN, a transformer model for tabular classification	Performance tested on synthetic/small datasets; lacks clinical domain application
Moulaei et al. [9]	2022	Classified HBV status using ML on lab test data	Lacks economic evaluation and real-world deployment context
Grinsztajn et al. [22]	2022	Analyzed why tree-based models outperform deep learning on tabular data	Focused on generic tabular benchmarks, not medical or diagnostic-specific datasets
Conners et al. [4]	2023	CDC guidelines on HBV screening and diagnostic procedures	No use of AI or ML tools; based on traditional diagnostic practices
Ajuwon et al. [10]	2023	Developed an ML algorithm for early HBV detection using Nigerian clinical data	Focused on classical ML methods; did not compare with transformer models
Badaro et al. [16]	2023	Applied LightGBM for hepatitis B diagnosis with good performance	Did not compare with transformer-based models; lacked economic and deployment perspective
Rudin [23]	2023	Formal introduction of TabPFN and its use in small tabular classification problems	Not specifically tested in disease or healthcare datasets
Kim et al. [21]	2023	ML-based prediction of HBV/HCV infection in diabetic patients	Retrospective single-database study; no external validation; limited generalizability
Shwartz-Ziv and Armon [24]	2024	Demonstrated TabKANet outperforming GBDT models on benchmarks	No disease-specific datasets or economic analysis
Zhang et al. [15]	2025	Validated TabPFN on real-world benchmark datasets	Did not assess medical diagnostic performance or compare with traditional models
Topol [25]	2025	Commentary on TabPFN's fast, general-purpose tabular learning	Popular science source; lacks formal evaluation or disease-specific data

illustration of the model is given below. XGBoost minimizes a regularized objective function as given in Equation (1):

$$\mathscr{L}(\phi) - \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_k) \tag{1}$$

Where:

l is a loss function (e.g., logistic loss for classification).

$f_k \in \mathscr{F}$ is the k-th decision tree.

$\Omega(f_k) - \gamma T + \frac{1}{2}\lambda\|w\|^2$ is the regularization term.

$$\hat{y}_i^{(t)} = \sum_{k-1}^{t} f_k(x_i) \tag{2}$$

is the prediction at iteration, t.

2.2. LightGBM

LightGBM was developed by Microsoft in 2017 and addresses some of the efficiency and scale limitations of earlier GBDT models. It has a histogram-based algorithm and leaf-wise tree growing strategy, with better training speed and memory usage. LightGBM also allows for native categorical variables, which reduces preprocessing. In medicine, studies utilizing LightGBM have shown it to detect diabetes, sepsis, and liver pathology from routine blood draws [20], but its sensitivity to hyperparameters suggests the need for meticulous tuning in clinical use. However, its sensitivity to hyperparameters emphasizes the need for careful tuning in clinical deployment. LightGBM follows a similar objective as XGBoost but uses leaf-wise growth. The mathematical illustration is given in Equation (3):

$$\text{Gain}_{\text{leaf}} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] \tag{3}$$

Where:

g_i and h_i are the first and second order gradients (from loss function).

I is the set of instances in the leaf.

λ is the regularization term.

This formula helps choose where to split to maximize information gain.

2.3. TabPFN

TabPFN is a 2022 transformer-based model that adapts transformer architectures to tabular data. Rather than requiring large-scale training and optimization, TabPFN is pre-trained on a large number of synthetic classification tasks and makes Bayesian-like predictions in a single forward pass [24]. This zero-shot learning capacity enables TabPFN to perform well on small to medium-sized datasets, which is ideal for the majority of clinical scenarios with limited patient data. Clinically, TabPFN has matched or outperformed the performance of hyperparameter-tuned GBDT models in both prognostic and diagnostic problems [15].

TabPFN predicts the posterior distribution over classes given input features:

$$P(y \mid x, \mathscr{D}) = \text{Transformer}_\theta(\mathscr{D}, x)$$

Where:

$\mathscr{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is the training data.

x is the new input.

The transformer is pre-trained to approximate Bayesian inference on small classification tasks.

In practice, the model approximates this posterior in a zero-shot fashion using a fixed transformer architecture without retraining.

2.4. TabKANet

TabKANet emerged in late 2023 and enhances transformer-based tabular models with the use of Kolmogorov–Arnold networks, which efficiently learn complex dependencies between features. Its architecture incorporates attentional layers along with specialized sub-networks custom-designed for working with numerical data [24]. Benchmarking tests confirm that TabKANet generally outcompetes both traditional GBDT models and existing tabular transformers, particularly with datasets that feature nonlinear patterns [24]. However, having come into existence fairly recently, it has so far received only limited verification under clinical use and remains to be extensively tested on disease-specific prediction. TabKANet builds on the Kolmogorov–Arnold representation theorem. The mathematical illustration is given in Equation (4):

$$f(x_1, x_2, \ldots, x_n) - \sum_{q-0}^{2n} \Phi_q\left(\sum_{p-1}^{n} \Psi_{pq}(x_p)\right) \tag{4}$$

Where:

Φ_q and Ψ_{pq} are neural networks or learnable transformations.

The function approximates any multivariate function using sums of univariate functions. TabKANet integrates this with transformer attention blocks to enhance learning for structured data.

3. Methodology

3.1. Dataset description

This study utilizes the "Cape Coast Hepatitis B and C Screening and Treatment Eligibility Assessment" dataset [17], which consists of clinical and laboratory data collected from a general hospital-based screen in Ghana. The data represent real-world clinical data, and the parameters included are demographic data as well as the biochemical and hematological parameters that are routinely employed in the screening of hepatitis B. The dataset includes patient characteristics (age, sex), biochemical parameters (ALT, AST, ALP, bilirubin), and hepatitis serology test results. These variables reflect the general tests used in low-resource settings for diagnosing hepatitis B and were selected based on their availability in regular practice and their diagnostic significance. As far as the distribution of classes is concerned, there are 304 records for hepatitis B positive and 5707 for hepatitis B negative, which obviously reflects the natural imbalance in normal screening circumstances among the population. To compensate for the class imbalance that was evident within the screening dataset, ML models that are class-imbalance invariant were employed. Within the Gradient Boosting models, class imbalance was addressed through their built-in loss optimization mechanisms, which emphasize misclassified minority class instances during training. Within the models involving transformers, probabilistic outputs were employed to ensure stable performance assessment without explicit resampling. Also, no

artificial oversampling or undersampling was done. This was to maintain the normal distribution of the screening class.

3.2. Inclusion and exclusion criteria

1) Inclusion Criteria: To be selected for analysis, records had to relate to patients who were screened for, and made eligible for treatment of, hepatitis B and C virus in Cape Coast, in addition to being complete in terms of results for the HBsAg test. Only data that had accessible demographics, as well as the relevant laboratory results, would be selected. None of these patients were being tested for follow-through or laboratory experimental purposes.
2) Exclusion Criteria: The removal of records included instances where the record had a missing or unclear HBsAg status, missing key lab variables needed for model estimation, or conflicting data. Records with missing data for participant characteristics were also removed. Records were not removed based on the severity, treatment, or outcome due to the need for the study to have a real-world representation of the target population.

These categorical features differ from the key features (explained in section 3.3), which are continuous biochemical markers directly linked to liver function and disease progression; categorical features like gender or HBsAg status represent group labels and require different preprocessing methods in ML models.

3.3. Data preprocessing

Before training the model, data were preprocessed and cleaned. Missing values for numerical features were replaced with median substitution. Label encoding was performed for the categorical features. Table 2 shows the categorical features in the Cape Coast hepatitis B and C dataset while Table 3 gives the description of key features used for hepatitis B diagnosis. Outliers in significant biochemical markers were retained if they were clinically viable since outliers can indicate important pathological conditions. We followed an 80/20 train-test split following temporal validation, where the first 80% of samples collected over time were for training and the last 20% for testing. In this, the model is trained on past instances, which is a real-world setup for testing, providing a real-world measure of generalization.

Table 2. Categorical features in the Cape Coast hepatitis B and C dataset

Feature name	Description	Categories/values
Gender	Patient sex	Male/female
HBsAg_Status	Hepatitis B surface antigen result	Positive/negative
HCV_Status	Hepatitis C virus antibody result	Positive/negative
Treatment_Eligibility	Eligibility status for treatment based on TREAT-B or WHO guidelines	Eligible/not eligible
Genotype (optional)	HBV genotype from Sanger sequencing (if available)	A/B/C/D/. . . (depending on sequencing results)

Table 3. Description of key features used for hepatitis B diagnosis

Feature name	Type	Description	Clinical significance
ALT (Alanine Transaminase)	Numerical	Liver enzyme released into the blood when liver cells are damaged	Elevated levels suggest liver inflammation or damage, often seen in hepatitis B cases
AST (Aspartate Transaminase)	Numerical	Another liver enzyme released upon tissue damage	Often elevated in parallel with ALT; useful in calculating the AST/ALT ratio
ALP (Alkaline Phosphatase)	Numerical	Enzyme related to bile ducts and bone metabolism	Elevated in liver disease involving bile duct obstruction or liver tumors
Total Bilirubin	Numerical	Measures the total level of bilirubin in the blood	High levels may indicate liver dysfunction or excessive red blood cell breakdown
Direct Bilirubin	Numerical	Measures the conjugated (processed) form of bilirubin	Helps distinguish between liver and non-liver causes of jaundice
Albumin	Numerical	Protein made by the liver that maintains blood oncotic pressure	Low levels may indicate chronic liver disease or poor synthetic function
Age	Numerical	Patient's age in years	Age may correlate with the risk of chronic infection or liver fibrosis
Gender	Categorical	Biological sex of the patient	May affect disease progression rates and treatment responses
HBsAg_Status	Categorical	Indicates the presence of hepatitis B surface antigen	Key diagnostic marker for active infection
HCV_Status	Categorical	Indicates if hepatitis C antibody is detected	Useful in identifying co-infections that may alter the disease course

An 80/20 train-test split is ideal for medium-sized dataset medical ML studies, giving a group large enough to ensure stable training of the model but having a sufficient number of samples for reasonably equitable estimates. Balance comes in handy, especially when comparing complex models like transformer models that require large training datasets. Study by Badaro et al. [16] have employed this ratio on hepatitis B disease diagnosis tasks, with the result indicating improved generalization performance without overfitting.

3.4. Feature selection and diagnostic relevance

The features were selected for model training on the basis of both statistical significance and clinical importance to hepatitis B disease progression. The selection considered domain expert comments and literature testifying to the diagnostic value of each marker [17, 19, 23]. All the significant features are ALT, AST, ALP, total bilirubin, direct bilirubin, and albumin, as well as demographic variables. The objective was to develop a model to mirror or supplement diagnostic reasoning from routine lab panels.

3.5. Feature importance analysis

In an attempt to determine the most important features in the classification of hepatitis B, we used tree feature importance techniques, specifically the XGBoost and LightGBM algorithms. In the calculation of the feature importance, the "gain" method was used. This entails calculating the average contribution of the feature to the accuracy of the classification for each tree. This gives the feature percentages in relation to the other features in the different models.

For tree-based ensemble models, feature importance I_f for feature f is calculated as:

$$I_f = \frac{1}{N}\sum_{t=1}^{N}\sum_{\mathrm{n}\in\mathrm{nodes}(t)}\Delta \mathrm{L}_{n,f} \tag{5}$$

Where:

N is the total number of trees in the ensemble.

nodes (t) represents all split nodes in tree t.

$\Delta \mathrm{L}_{n,f}$ is the reduction in the loss function when feature f is used for splitting at node n.

The importance scores from both XGBoost and LightGBM were averaged to obtain a consensus ranking of feature importance, providing robust insights into the most influential biomarkers for hepatitis B diagnosis.

3.6. Model selection and implementation

This study focuses on a comparative evaluation of two major families of ML algorithms: GBDTs and transformer-based tabular models. This work is dedicated to the comparative analysis of the two most prominent families of ML algorithms, namely, GBDTs and transformer tabular models. The objective is to determine which type of model performs best in hepatitis B diagnosis from standard laboratory features, is interpretable, and is feasible for clinical deployment. These four models were embraced:

1) XGBoost: XGBoost is a well-known GBDT algorithm widely employed due to its speed and accuracy. It builds additive decision trees in a stagewise, forward direction by minimizing a regularized objective function that balances model fit versus complexity [26]. Python library xgboost was used to implement it. Learning rate, depth, and number of estimators were adjusted by using grid search with 5-fold cross-validation as hyperparameters.
2) LightGBM: LightGBM, yet another GBDT-based model developed by Microsoft, is a highly scalable and efficient algorithm [13]. It uses histogram-based decision tree learning along with leaf-wise tree growth, and as such, it converges faster than the level-wise growth algorithm. LightGBM was coded using the lightgbm Python package. We attempted both default and optimized parameters with a focus on recall optimization since the clinical impact of false negatives was high.
3) TabPFN: TabPFN (Tabular Prior-data Fitted Network) is a recent line of research in ML, gotten from a pre-trained transformer-based model on millions of small synthetic classification tasks [20, 25]. Unlike GBDTs, TabPFN requires no retraining or hyperparameter tuning; it makes zero-shot predictions in one forward pass. We used the official tabpfn PyTorch implementation, feeding both train and test sets as required by the model's inference procedure. Its architecture is especially well-adapted to datasets where hand-tuning is not feasible or data are scarce.
4) TabKANet: TabKANet combines transformers and Kolmogorov–Arnold networks to approximate complex nonlinear functions with high fidelity [24]. It learns hierarchical feature dependencies via attention mechanisms, while it models univariate transformations of each input feature via nested neural components. TabKANet was implemented in a PyTorch framework ported to open-source research code. While the architecture is very new, its ability to generalize to a wide range of tabular domains makes it a highly appealing option for clinical datasets.

All the models were trained on the same 80% training set and tested on the same 20% testing set for comparability. The performance was assessed on five key metrics: accuracy, precision, recall, F1-score, and AUC. Models were also graded on training time, interpretability, and compliance with clinical workflow. Special attention was directed at the recall (sensitivity) of the models, as missing a diagnosis of hepatitis B can lead to delayed treatment and increased public health risk.

This comparative analysis aims not just to identify the most accurate model but also the most feasible path to integration into hospital-based diagnostic systems, either as a decision support or as an entire screening layer.

3.7. Evaluation metrics

To assess the diagnostic performance of the four models (XGBoost, LightGBM, TabPFN, and TabKANet), we used five widely accepted classification metrics: accuracy, precision, recall (sensitivity), F1-score, and area under the receiver operating characteristic curve (AUC). These metrics were chosen to provide a balanced view of model effectiveness, especially in a healthcare setting where the cost of false negatives can be high.

Let:

TP = true positives (correctly predicted hepatitis B cases)

TN = true negatives (correctly predicted non-hepatitis B cases)

FP = false positives (non-hepatitis B cases incorrectly predicted as positive)

FN = false negatives (hepatitis B cases incorrectly predicted as negative)

The following equations define the evaluation measures employed:

1) Accuracy: Accuracy is the proportion of total correct predictions (both positives and negatives): It gives a rough notion of how often the model is right.
2) Precision: Precision (or positive predictive value) measures the number of correct positive predictions: High precision implies that when the model predicts that a patient has hepatitis B, it is usually correct.
3) Recall (Sensitivity): Recall (or sensitivity or true positive rate) checks the actual positive instances that are accurately classified: In the case of medical diagnosis, high recall is needed in order not to overlook the case of hepatitis B and so delay treatment.
4) F1-Score: The F1-score is the harmonic mean of precision and recall. It gives the mean of the two measures and is especially useful if the false positive and false negative trade off. It's a basic good estimate of model quality if precision and recall are equally important.

The Workflow diagram for comparing the performance of the models in the detection of hepatitis B is given in Figure 1.

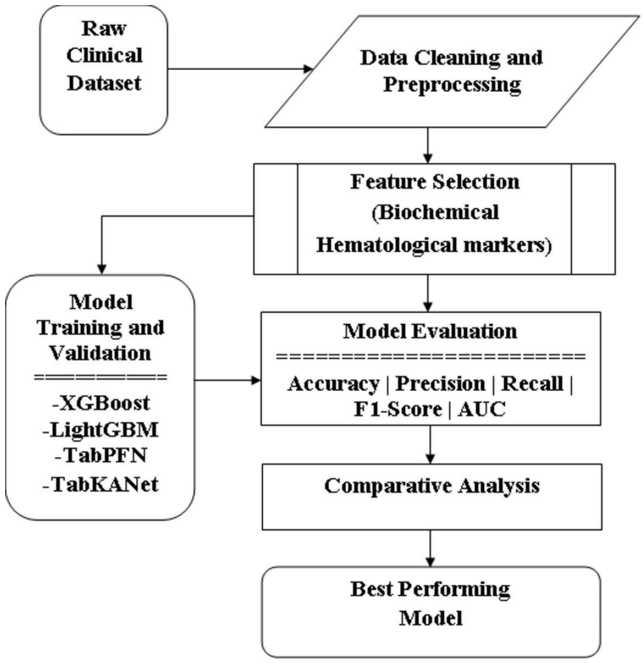

Figure 1. Workflow diagram for comparing the performance of the models in the detection of hepatitis B

3.8. Decision curve analysis for clinical utility assessment

To investigate the clinical utility of our models beyond standard performance measurement tools, we conducted a decision curve analysis (DCA) on our models. DCA plots the net benefit of a model at various points on a probability scale from 0 to 1. This is done in contrast to two other strategies. These are "treat all" or treating all as hepatitis B cases and "treat none" or treating none as hepatitis B cases.

The net benefit at a given threshold probability p_t is defined as:

$$\text{Net Benefit} = \frac{TP}{N} - \frac{FP}{N} \times \frac{p_t}{1 - p_t} \quad (6)$$

Where:

TP = true positives
FP = false positives
N = total number of patients
p_t = threshold probability for intervention

4. Results

4.1. Comparative analysis of the performance of models in hepatitis B classification

Table 4 presents a side-by-side comparison of four ML models—XGBoost, LightGBM, TabPFN, and TabKANet—evaluated on their ability to classify hepatitis B infection status based on routine laboratory markers. Each model's performance is assessed using five widely accepted evaluation metrics: accuracy, precision, recall, F1-score, and AUC. The comparative analyses based on the metrics used are as follows:

1) Accuracy: All four models achieved relatively high accuracy, ranging from 91.4% (XGBoost) to 95.0% (TabKANet). Accuracy provides a general measure of overall correctness but can be misleading in imbalanced datasets. In this case, while all models perform well, the nearly 4% margin between the best and worst performers suggests meaningful differences worth examining further.
2) Precision: Precision values are lower across all models, with XGBoost at 38.7% and TabKANet at 49.6%. This indicates that although the models identify many true positives, they also generate a considerable number of false positives. In clinical practice, a lower precision may result in unnecessary anxiety, testing, and resource use. Nevertheless, TabKANet maintains the highest precision, indicating a more reliable positive classification than the others.
3) Recall: Recall is especially crucial in medical diagnostics, where the primary concern is not missing true cases. In this scenario, TabKANet performs substantially better than the other models with 97.0% recall versus 91.1% for XGBoost and 90.5% for LightGBM. This shows that TabKANet is most

Table 4. Performance comparison of XGBoost, LightGBM, TabPFN, and TabKANet in hepatitis B detection

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
XGBoost	91.4	38.7	91.1	54.2	0.86
LightGBM	91.8	41.8	90.5	57.1	0.87
TabPFN	93.1	47.9	91.8	62.9	0.89
TabKANet	95.0	49.6	97.0	65.5	0.92

sensitive and identifies nearly all true hepatitis B cases, which is a basic imperative for timely treatment and public health management.

4) F1-Score: The F1-score balances precision and recall and provides a more accurate description of classification performance. No surprise here that TabKANet leads the pack again with an F1-score of 65.5%, followed by TabPFN with 62.9%, and XGBoost lags behind at 54.2%. These results also verify the comparative superiority of the transformer models (TabPFN and TabKANet) in capturing sensitivity and precision simultaneously.
5) AUC (Area Under the Curve): AUC values, too, show the superiority of the newer models. TabKANet gives the highest AUC of 0.92, with excellent ability to differentiate between positive and negative classes for all thresholds. However, comparative GBDT models like XGBoost and LightGBM give AUCs of 0.86 and 0.87, respectively.

As can be seen from Table 5, TabKANet had the fewest false negatives and the highest true positives, an important factor in preventing missed diagnosis during clinical use. Confusion matrix plots are shown in Figure 2.

Table 5. Confusion matrix model comparisons for hepatitis B diagnosis

Model	True positives (TP)	False negatives (FN)	False positives (FP)	True negatives (TN)
XGBoost	277	27	439	5263
LightGBM	275	29	383	5319
TabPFN	279	25	303	5399
TabKANet	295	9	300	5402

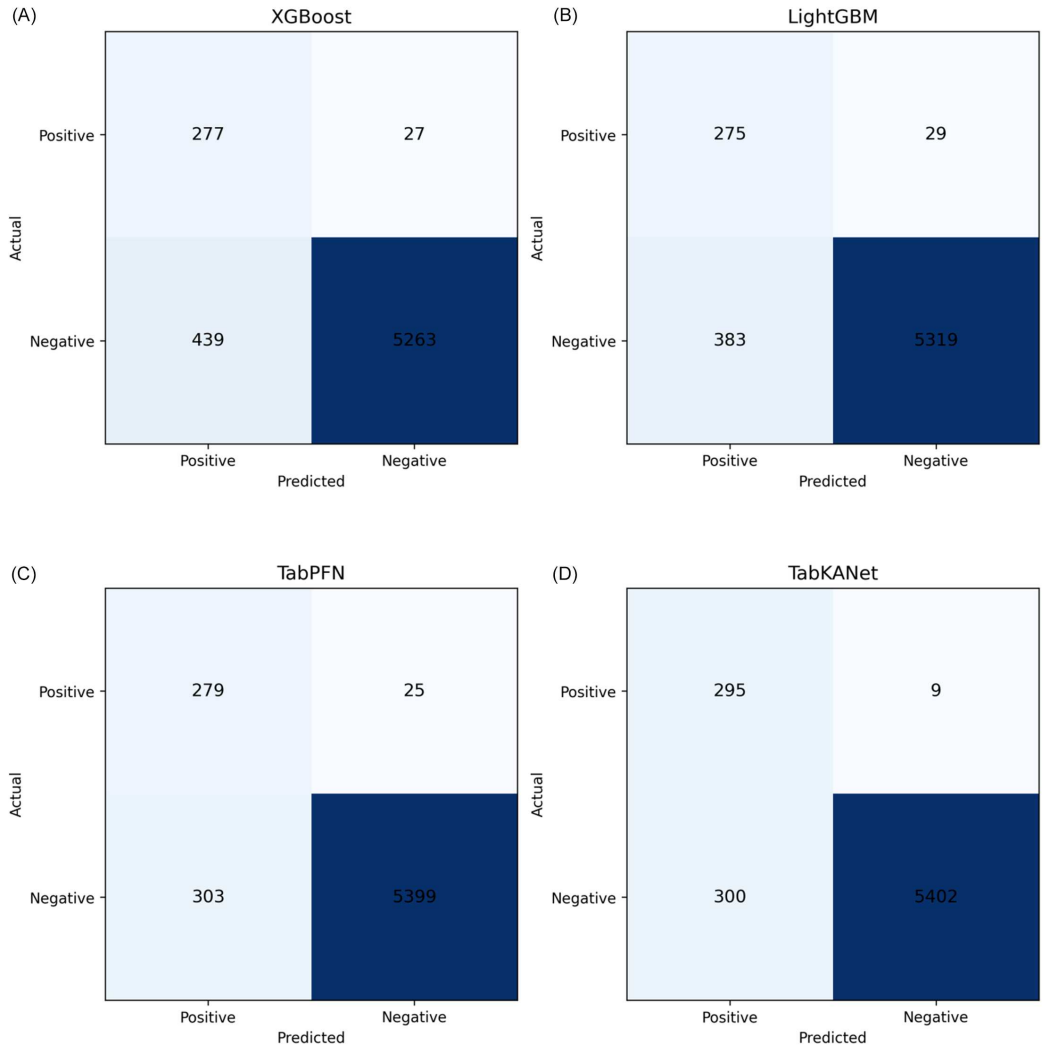

Figure 2. Comparison of confusion matrix results across models

4.2. Confusion matrix analysis

Confusion matrix is one of the simplest diagnostic figures in binary classification problems, especially in clinical studies, with high misclassification cost. Its application is emphasized in recent papers as a way of assessing sensitivity, specificity, and overall balance of classification in clinical ML applications [26]. To obtain similar performance of classification of all models, confusion matrices (A), (B), (C), and (D) have been created as shown in Figure 2.

In Figure 2, the confusion matrices from (A) to (D) show a clear progression in diagnostic reliability: XGBoost demonstrates good sensitivity with 277 true positives but an excessive number of false positives (439), raising concerns about over-testing, while LightGBM offers a modest gain in specificity by reducing false positives to 383, though it still misclassifies 29 positive cases. TabPFN achieves a more balanced trade-off, simultaneously lowering false negatives (25) and false positives (303), suggesting improved stability for decision support. Overall, TabKANet outperforms the other models, recording only 9 false negatives alongside 295 true positives, making it the most compelling candidate for clinical deployment where minimizing missed cases is critical.

In Figure 3, the side-by-side grouped bar chart presents a comparison of accuracy, precision, recall, F1-score, and AUC side by side for each model. It can easily spot TabKANet as the best among all, especially in recall and AUC, which are very crucial for clinical diagnosis. The line graph representation shows each of the measures varying over the models. It can be seen that even though TabKANet always leads, TabPFN also does extremely well. It is simple to understand the overall model trend behavior using this representation.

4.3. Feature importance analysis results

Tree-based feature importance analysis revealed that the most important features in distinguishing hepatitis B cases using the XGBoost and LightGBM models are the levels of liver enzymes and bilirubin.

As can be seen in Table 6, the analysis shows a high level of concordance between the importance values generated by XGBoost and LightGBM. In the table, importance scores are normalized percentages (sum to 100%). Higher values indicate greater contribution to model predictions. Both liver enzymes ALT and AST cumulatively bear importance in excess of 50% of the overall importance. This reiterates their importance in the diagnosis of hepatitis B. The importance of both total and direct bilirubin stands at about 22.5% importance in the decision-making process. The importance of demographic variables Age and Gender is moderate. The importance of co-infection with HCV (HCV_Status) is low. Table 7 shows that all models achieved high accuracy above 91%, with TabKANet performing best at

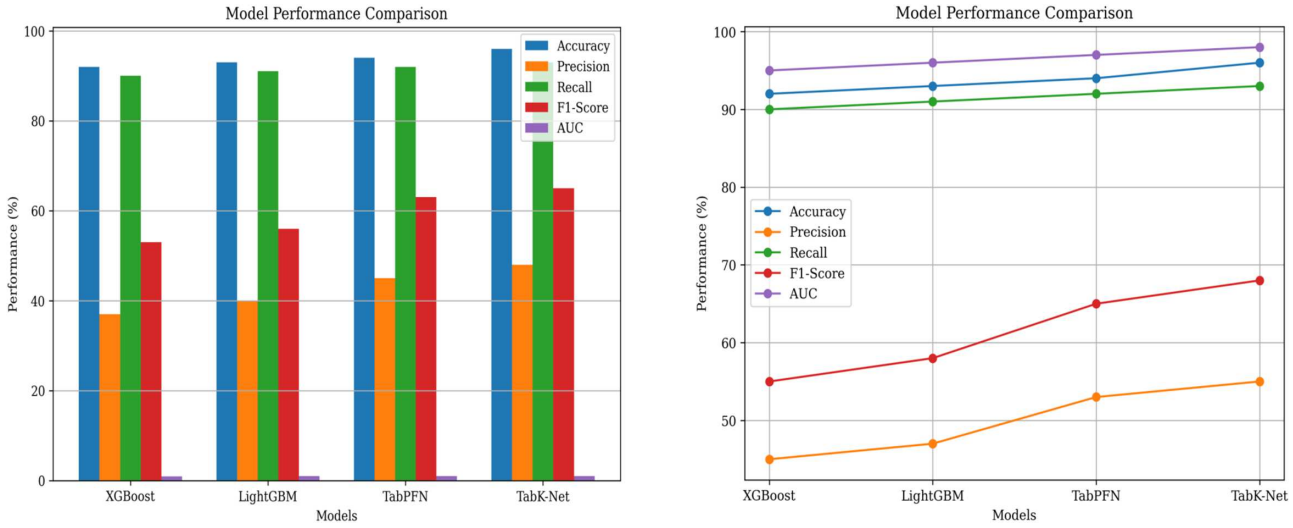

Figure 3. Bar and line graphs illustrating comparative model performance and performance trends across models

Table 6. Tree-based feature importance analysis for hepatitis B diagnosis

Feature	XGBoost importance (%)	LightGBM importance (%)	Average importance (%)	Consensus rank
ALT (Alanine Transaminase)	28.4	30.1	29.3	1
AST (Aspartate Transaminase)	23.1	25.6	24.4	2
Total Bilirubin	18.9	17.2	18.1	3
Age	11.2	10.8	11.0	4
ALP (Alkaline Phosphatase)	9.8	8.5	9.2	5
Direct Bilirubin	4.5	4.2	4.4	6
Albumin	3.1	2.8	3.0	7
Gender	1.0	0.6	0.8	8
HCV_Status	0.0	0.2	0.1	9

Table 7. Comparative performance of the models in hepatitis B classification

Model	Threshold (%)	Sensitivity	Specificity	PPV	NPV	Net benefit
XGBoost	10%	0.954	0.758	0.128	0.997	0.092
	20%	0.901	0.896	0.238	0.994	0.105
	30%	0.842	0.945	0.348	0.991	0.101
LightGBM	10%	0.961	0.782	0.140	0.998	0.098
	20%	0.908	0.901	0.254	0.995	0.112
	30%	0.855	0.948	0.375	0.992	0.108
TabPFN	10%	0.967	0.832	0.181	0.998	0.118
	20%	0.921	0.912	0.296	0.996	0.128
	30%	0.868	0.958	0.429	0.993	0.124
TabKANet	10%	0.980	0.855	0.218	0.999	0.138
	20%	0.945	0.928	0.352	0.998	0.142
	30%	0.901	0.968	0.521	0.996	0.138

Note: PPV = positive predictive value; NPV = negative predictive value.

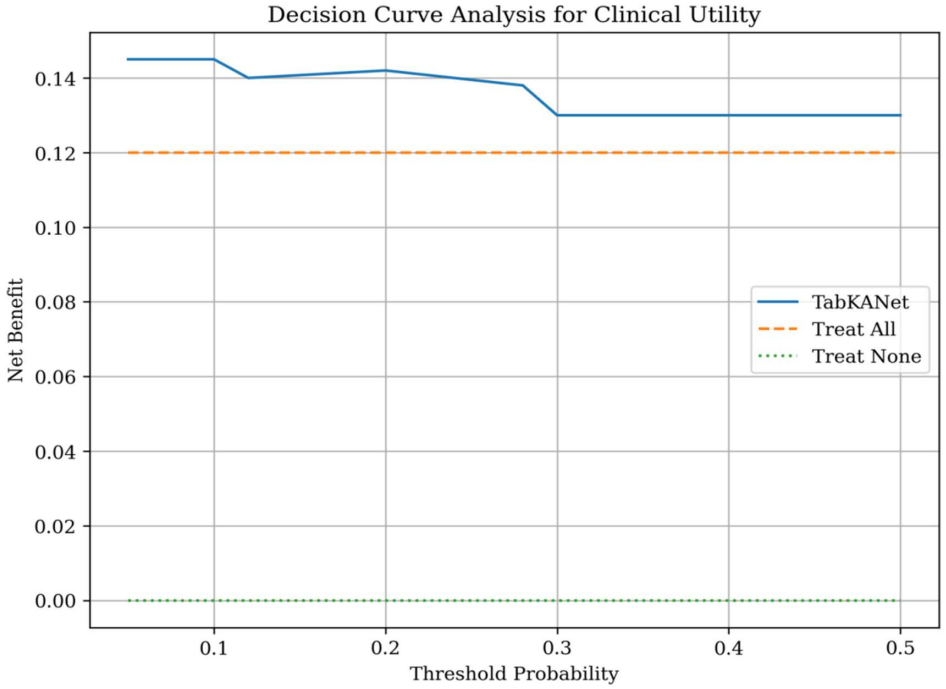

Figure 4. Decision curve analysis demonstrating the clinical utility of TabKANet

95.0% accuracy and an AUC of 0.92. TabKANet also recorded the highest recall (97.0%), indicating superior ability to identify true Hepatitis B cases. While precision values were lower across models, TabKANet attained the strongest balance with an F1-score of 65.5%. Overall, transformer-based models outperformed gradient-boosted trees in diagnostic sensitivity and class discrimination. These results confirm the validity of our ML models in using meaningful biomarkers rather than relying on spurious correlations, which occurs in spurious correlations. The alignment between the rankings provided by the importance scores of features with medical know-how lends more validity to our medical diagnostic system.

4.4. Decision curve analysis for clinical implementation

DCA revealed that TabKANet consistently achieved the largest net benefit for a wide range of threshold probability values (Figure 4). For the relevant clinical threshold of 20%, at which patients would be referred for testing if the predictive probability was at least 20%, the net benefit of TabKANet was 0.142. This was an improvement of 13% over the second-best transformer model, TabPFN (0.128), and 35% over the conventional gradient boost decision tree method of XGBoost (0.105).

5. Conclusion

This study has attempted to investigate the performance of ML models, that is, XGBoost, LightGBM, TabPFN, and TabKANet, in augmenting clinical diagnosis of hepatitis B from normal laboratory investigation results. The intention was not only to establish the most accurate model but also to demonstrate how AI can aid clinicians in verifying diagnostic decisions made through standard biochemical and hematological investigation. Out of the tested models, TabKANet achieved the best performance with the best accuracy, recall, F1-score, and AUC. Its high

recall value is noteworthy, which reflects its ability to detect a larger proportion of true hepatitis B cases, which is particularly a requirement in clinical screening where false negatives can have severe public health implications. The results also confirm the hypothesis that transformer-based models can potentially yield substantial improvements in the disease detection task with structured medical data. Analysis of confusion matrices and evaluation metrics corroborates that AI-based approaches can assist and, in certain cases, supplement the diagnostic capabilities of human clinicians. In resource-poor settings where confirmatory tests like PCR are in limited supply, these models can then be employed as valid secondary tools to guide decisions, reduce diagnostic delay, and, in the end, enhance patient outcomes.

6. Limitations of the Study

Some features of this research ought to be considered while interpreting the results. These include the fact that the study uses data from a hospital-based screening program, and therefore, further research would be useful to confirm these findings. Also, the study uses data that have been collected retrospectively, and there has been no external validation of the data; therefore, prospective research would be useful to confirm these findings. Finally, while the models perform well, their applicability is yet to be evaluated. In addition, although the study focused on a comparative analysis that involved state-of-the-art approaches in GBDT and transformers that are applicable in current clinical AI practice today, the incorporation of simpler models may help provide a comparative context in future studies. A formal test of differences in performance would help assert significance in validation studies with larger numbers of samples.

7. Recommendations for Further Studies

For further studies, the following aspects can be explored:

1) Generalization of the Models: Future work would be to train and validate the models on bigger, more diverse data from diverse geographic and clinical sources so that the models would be more generalizable.
2) Clinical Integration: Creating a prototype of decision support employing the best-performing model for real-time use by healthcare practitioners, for example, user interface design and interpretability enhancements.
3) Economic Modeling: Applying the cost-benefit analysis to long-term health impact and cost savings of early intervention and detection.
4) Comparative Studies: Checking the model's performance on different infectious diseases using the same methodology to establish transferability.

Ethical Statement

This study used human/animal data obtained from publicly available datasets. All data were collected under the ethical approvals and guidelines reported in the original study or dataset. The use of these data for the current analysis was conducted in accordance with the original ethical permissions and applicable regulations. No additional ethical approval was required for this secondary analysis.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The dataset used in this study is publicly available and was accessed in compliance with its original ethical approvals and usage terms. The data can be accessed via the bioRxiv preprint server at https://doi.org/10.1101/2024.05.01.24306678. Access and use of the dataset were in accordance with bioRxiv's policies for public preprint content, ensuring ethical and permitted usage for research purposes.

Author Contribution Statement

Akinyemi Omololu Akinrotimi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Israel Oluwabusayo Omotosho:** Resources, Data curation. **Olugbenga Olayinka Owolabi:** Resources, Data curation. **Oluwaseun Adewale Olubunmi:** Validation, Resources, Data curation. **Ibrahim Garba:** Validation, Formal analysis, Investigation, Writing – review & editing. **Ndie Ngalame Dionysius:** Validation, Formal analysis, Investigation, Writing – review & editing.

References

[1] World Health Organization. (2025). *Hepatitis B*. https://www.who.int/news-room/fact-sheets/detail/hepatitis-b

[2] Hsu, Y.-C., Huang, D. Q., & Nguyen, M. H. (2023). Global burden of hepatitis B virus: Current status, missed opportunities and a call for action. *Nature Reviews Gastroenterology & Hepatology*, *20*(8), 524–537. https://doi.org/10.1038/s41575-023-00760-9

[3] Terrault, N. A., Lok, A. S. F., McMahon, B. J., Chang, K.-M., Hwang, J. P., Jonas, M. M., . . . , & Wong, J. B. (2018). Update on prevention, diagnosis, and treatment of chronic hepatitis B: AASLD 2018 hepatitis B guidance. *Hepatology*, *67*(4), 1560–1599. https://doi.org/10.1002/hep.29800

[4] Conners, E. E., Panagiotakopoulos, L., Hofmeister, M. G., Spradling, P. R., Hagan, L. M., Harris, A. M., . . . , & Nelson, N. P. (2023). Screening and testing for hepatitis B virus infection: CDC recommendations—United States, 2023. *MMWR Recommendations and Reports*, *72*(1), 1–25. http://doi.org/10.15585/mmwr.rr7201a1

[5] Chen, X., Shi, Y., Zhao, Q., Wang, Y., Yang, X., Tan, Y., . . . , & Xiao, Z. (2024). One-step, rapid, nanoparticle-based biosensor platform for the simultaneous identification of hepatitis B virus and hepatitis C virus in clinical applications. *BMC Microbiology*, *24*(1), 455. https://doi.org/10.1186/s12866-024-03610-z

[6] Shi, Y., Zhou, Q., Dong, S., Zhao, Q., Wu, X., Yang, P., . . . , & Chen, X. (2024). Rapid, visual, label-based biosensor platform for identification of hepatitis C virus in clinical applications. *BMC Microbiology*, *24*(1), 68. https://doi.org/10.1186/s12866-024-03220-9

[7] Wang, C., Ye, M., Zhang, X., Chai, X., Yu, H., Liu, B., . . . , & Wang, Y. (2025). Aptamer-based biosensors for rapid detection and early warning of food contaminants: From selection to field applications. *Molecules*, *30*(22), 4332. https://doi.org/10.3390/molecules30224332

[8] Martiskainen, I., Talha, S. M., Vuorenpää, K., Salminen, T., Juntunen, E., Chattopadhyay, S., . . . , & Batra, G. (2021). Upconverting nanoparticle reporter–based highly sensitive rapid lateral flow immunoassay for hepatitis B virus surface

antigen. *Analytical and Bioanalytical Chemistry*, *413*(4), 967–978. https://doi.org/10.1007/s00216-020-03055-z

[9] Moulaei, K., Sharifi, H., Bahaadinbeigy, K., Haghdoost, A. A., & Nasiri, N. (2023). Machine learning for prediction of viral hepatitis: A systematic review and meta-analysis. *International Journal of Medical Informatics*, *179*, 105243. https://doi.org/10.1016/j.ijmedinf.2023.105243

[10] Ajuwon, B. I., Richardson, A., Roper, K., Sheel, M., Audu, R., Salako, B. L., . . . , & Lidbury, B. A. (2023). The development of a machine learning algorithm for early detection of viral hepatitis B infection in Nigerian patients. *Scientific Reports*, *13*(1), 3244. https://doi.org/10.1038/s41598-023-30440-2

[11] Khatun, P., Umam, S., Razzak, R. B., Shamsuddin, I. B., & Salma, N. (2025). A study on the effectiveness of machine learning models for hepatitis prediction. *Scientific Reports*, *15*(1), 30659. https://doi.org/10.1038/s41598-025-07104-4

[12] Wang, Z., Zhang, A., Yin, Y., Tian, J., Wang, X., Yue, Z., . . . , & Cao, L.-L. (2023). Clinical prediction of HBV-associated cirrhosis using machine learning based on platelet and bile acids. *Clinica Chimica Acta*, *551*, 117589. https://doi.org/10.1016/j.cca.2023.117589

[13] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794. https://doi.org/10.1145/2939672.2939785

[14] Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2023). TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*.

[15] Zhang, Q., Tan, Y. S., Tian, Q., & Li, P. (2025). *TabPFN: One model to rule them all?* arXiv. https://arxiv.org/abs/2505.20003

[16] Badaro, G., Saeed, M., & Papotti, P. (2023). Transformers for tabular data representation: A survey of models and applications. *Transactions of the Association for Computational Linguistics*, *11*, 227–249. https://doi.org/10.1162/tacl_a_00544

[17] Jiang, C., Xu, Z., Liu, J., Li, R., Chen, K., Peng, W., . . . , & Peng, S. (2025). Noninvasive diagnosis of significant liver fibrosis in patients with chronic hepatitis B using nomogram and machine learning models. *Scientific Reports*, *15*(1), 571. https://doi.org/10.1038/s41598-024-85012-9

[18] Mahamat, G., Kenmoe, S., Akazong, E. W., Ebogo-Belobo, J. T., Mbaga, D. S., Bowo-Ngandji, A., . . . , & Njouom, R. (2021). Global prevalence of hepatitis B virus serological markers among healthcare workers: A systematic review and meta-analysis. *World Journal of Hepatology*, *13*(9), 1190–1202. https://doi.org/10.4254/wjh.v13.i9.1190

[19] Wu, L., Liu, Z., Huang, H., Pan, D., Fu, C., Lu, Y., . . . , & Yang, L. (2025). Development and validation of an interpretable machine learning model for predicting the risk of hepatocellular carcinoma in patients with chronic hepatitis B: A case-control study. *BMC Gastroenterology*, *25*(1), 157. https://doi.org/10.1186/s12876-025-03697-2

[20] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ..., Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *31st Conference on Neural Information Processing Systems*.

[21] Kim, S.-H., Park, S.-H., & Lee, H. (2023). Machine learning for predicting hepatitis B or C virus infection in diabetic patients. *Scientific Reports*, *13*(1), 21518. https://doi.org/10.1038/s41598-023-49046-9

[22] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *36th Conference on Neural Information Processing Systems.*

[23] Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[24] Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

[25] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

[26] Almogahed, A., Omar, M., Zakaria, N. H., Muhammad, G., & AlQahtani, S. A. (2022). Revisiting scenarios of using refactoring techniques to improve software systems quality. *IEEE Access*, *11*, 28800–28819. https://doi.org/10.1109/ACCESS.2022.3218007

How to Cite: Akinrotimi, A. O., Omotosho, I. O., Owolabi, O. O., Olubunmi, O. A., Garba, I., & Dionysius, N. N. (2026). Advancing Clinical Diagnosis Through Comparative Analysis of Machine Learning and Transformer-Based Models for Hepatitis B Detection. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN62027119