Research Article

Reliable and Calibrated Uncertainty Quantification for COVID-19 using Probabilistic Machine Learning Models

Rana Tarannum Ziyauddin Shaikh[1], Dheva Rajan S[2], Gargi Tyagi[3]

[1]Research Scholar, Department of Mathematics and Statistics; Banasthali Vidyapith ; Tonk, India. Email: rana.ansari@utas.edu.om
[2]Co- Supervisor, Department of Mathematics and Statistics; Banasthali Vidyapith; Tonk, India. Email: Dheva.Rajan@utas.edu.om
[3]Supervisor, Assistant Professor, Department of Mathematics and Statistics; Banasthali Vidyapith; Tonk, India. Email: tyagi.gargi@gmail.com

Abstract

Standard regression-based models have limited predictive capability because they often overlook nonlinear interactions among variables, including Peripheral Oxygen Saturation, metabolic markers, illness duration, and the need for Intensive Care Unit admission. The purpose of this study is to compare the generalization behavior and predictive performance of three supervised learning models, such as Random Forest (RF), Multinomial Logistic Regression(MLR), and Logistic Regression(LR), in the classification of Covid patient outcomes (Expired vs. Cured) and identify critical clinical determinants of mortality. In all the evaluations RF model has shown better performance in test AUC (≈0.95), accuracy (≈0.88), sensitivity (≈0.78), specificity (≈0.96), and Kappa (≈0.76). On the other hand, LR and MLR have a lower performance test AUC (≈0.65), accuracy (≈0.58), sensitivity (≈0.22), specificity (≈0.84), and Kappa (≈0.07). The 10-fold cross-validation endorsed the dominance of RF. This research highlights the necessity for data-driven models to pinpoint high-risk COVID-19 patients. It uses clinical data from COVID-19 patients to compare the performance of RF, MLR, and LR classifiers for cured vs expired COVID-19 cases.

Keywords: Regression, Logistic, Multinomial, Random Forest, Machine Learning, Clinical Modeling.

1. Introduction

For machine learning (ML), we will use baseline deterministic models, such as Random Forest (RF), and Artificial Neural Networks (ANN). For probabilistic uncertainty models, Bayesian Ridge Regression (BRR), quantile regression forest, Natural Gradient Boosting (NGboost), and conformal prediction on ANN/RF. The Bayesian Ridge Regression (BRR) model uses a Gaussian prior on the coefficients and a ridge penalty. The ridge logistics are implemented via the glmnet package with (α=0). Uncertainty has been estimated via bootstrap. Stratified CV k-folds (k=5) have been performed to prevent Area Under the Curve (AUC) folds in a single-class setting. The COVID-19 pandemic has hastened the use of ML models in healthcare to predict outcomes, triage patients, and allocate resources. It has been shown to be highly predictive, achieving high accuracy and AUC across many studies. This promises well for using ML as a decision-support tool in clinical settings. However, its high discrimination does not ensure accurate probability estimation, and this latter point is extremely important in safety-critical clinical scenarios (Niculescu-Mizil & Caruana, 2005; Van Calster et al., 2019). Clinicians use risk and confidence rather than binary outputs, so it is as important to calibrate the model as to get it correct.

The effectiveness of Probabilistic and Bayesian models are often limited by the need for parametric assumptions (Gelman et al., 2013). Madhan et al., (2025) used ML analysis for Text similarity assessments. Harmonic mean has been utilized for the similarity detection with Particle Swarm (Madhan, 2024). Conformal prediction has also been proposed in recent years as an alternative to addressing the problem of predictive uncertainty in ML models, without any assumption of a particular distribution, and with a guarantee of uncertainty in ML models in finite samples (Angelopoulos & Bates, 2023; Vovk et al., 2005). Although conformal prediction has shown significant theoretical promise, it has not achieved sufficient traction in healthcare ML. The advantages of using this method over deterministic and Bayesian ML models have not been fully explored.

The study is based on a real-world, clinically curated dataset from a single tertiary care hospital, enabling controlled internal validation. However, it is limited to immediate cross-institutional generalization using single-center data. The Bayesian models are deliberately used as interpretable probabilistic baselines rather than optimized discriminative learners. Conformal prediction guarantees are held under the standard exchangeability

assumption, which is reasonable for retrospective clinical cohorts. The study focuses on probability reliability rather than on optimizing decision thresholds. Furthermore, most COVID-19 ML studies test models in isolation or report only accuracy-based measures, with little or no reporting of calibration scores, interval coverage, and uncertainty sharpness (Gneiting & Raftery, 2007). These fragmented assessments of ML models hardly contribute to clinical trust. To address these concerns, this paper conducts a comprehensive evaluation of deterministic, probabilistic, and conformal ML models with respect to discrimination, calibration, and the reliability of their uncertainty estimates.

Research Objectives

1. To compare the accuracy and reliability of ML models.
2. To quantify and validate the uncertainty using interval-based measures.

2. Review of Literature

The current approach to this work relies on probabilistic ML models that incorporate uncertainty to describe probability distributions. Khodabakhshian et al., (2023), in their systematic review and comparison of Deterministic and Probabilistic Risk Management Approaches in Construction Projects have done so. Xiong et al., (2024) who have elaborated the calibration of subjective data biases. They have interpreted the model's predictive probabilities as ML-based predictions. Another work on adaptive importance sampling for Bayesian probability prediction has been conducted by Huang et al., (2023). Xiong et al., (2022) have reported a comparison of several ML techniques for predicting COVID-19 severity. It was discovered that RF performed better than SVM in terms of accuracy and AUC. Luo et al., (2021) have employed the Long Term Short Memory (LSTM) and Extreme Gradient Boosting (XGboost) model to predict its results, and the performance was verified by applying accuracy and MSE respectively. Devi et al. (2023) presented an IoT blockchain automation model for web applications. N Uday et al., (2024) presented an analysis of KNN and SVM algorithms in predicting human emotions. Maad et al., (2022) did the review work on deep learning (DL) image analysis. In their work, they have demonstrated that the clinical dataset can be improved by employing AI techniques in DL.

Singh et al., (2024) proposed Bayesian binary search model. They compared supervised and unsupervised ML models and showed that Bayesian supervised methods surpass others. Alzraiee & Niswonger, (2024) proposed probabilistic ways to train ML Models for noisy data. Ajirak et al., (2025) has done work on Uncertainty Quantification in Probabilistic ML Models. They have shown that the provided insight into the impact of uncertainty estimation on model reliability. Wang, (2024) explored probabilistic models for semi-supervised ML models and proposed the superiority of that. Ghosh & Barnett, (2025) compared Bayesian NN vs. Mixture Density Networks. Tyralis & Papacharalampous, (2024), review ML models of uncertainty. The modeling of observational error does not yet seem to keep up with. Dew et al., (2024) addressed the challenges of discovery of heterogeneity, flexible representation of nonlinearities and dynamics, very high-dimensional and unstructured data, and missingness often by fusion across disparate data sources. Kapoor, (2024) has worked on probabilities and AI.

They have discussed probabilistic ML modelling with real-world applications. In this, it is mentioned that most ML models, especially probability-oriented models, evaluate performance using "accuracy" and sensitivity measures. Nimy et al., (2023) utilized probabilistic ML models in students' risk assessment. Marvasti et al., (2021) proposed a framework for probabilistic ML. In this, they addressed the maximum probability theorem as a consequence of Kolmogorov's Probability theory, which is used to understand the model's behaviour. Sprangers et al., (2021) has done work on Probabilistic Gradient Boosting for Large-Scale Probabilistic Regression models.

2.1 Research Gap

Prior studies used deterministic performance metrics such as accuracy and sensitivity. This leads to strong discrimination, which in turn yields more reliable predictions. Probability calibration, uncertainty, and reliability are

rarely evaluated across different scenarios, and consistency is not assessed. Probabilistic models are seldom systematically compared with deterministic models under the same experimental work. Interval-based uncertainty measures such as Prediction Interval Coverage Probability (PICP) and Mean prediction interval (MPIW) are entirely absent in clinical ML studies.

3. Methodology

3.1 About Data

After cleaning, the data has 526 rows, 34 columns, and the selected target column is Discharge Status, which is predicted using 10 feature variables. It has a total of 173 expired and 353 cured cases. In this, the ex-ventilator, most critical, and DAMA cases are also considered expired for this analysis. The four models, RF, KNN, ANN, and SVM, were rigorously evaluated using a dataset collected from the Sara Multispecialty Hospital in Maharashtra, India. The Train: test split ratio is 80:20. The top numeric feature distributions are shown in Figure 1.

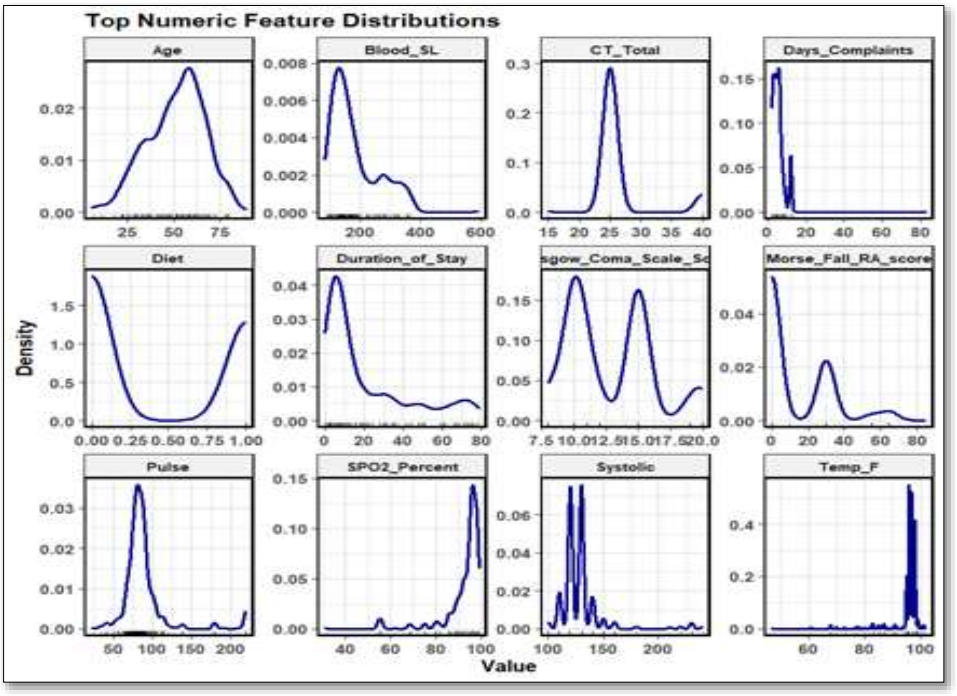


Fig. 1 Distribution of the Top numeric features in the data

Figure 1 is not just exploratory but also diagnostic. The distributions show that the features are either non-normal or skewed. Age shows a strong demographic risk relevance between 45 and 60 years. Blood sugar has a long tail, extending beyond 300-400 mg/dL, which supports robust models. Days of complaint show high diagnostic consistency with a sparse tail, but CT Total shows normality. Diet is a non-numeric variable, so the model treats it as a factor. The relationship between the Morse Fall Risk Assessment Score and the outcome is nonlinear, based on a multimodal curve. Oxygen Saturation (SPO2) percent has a high mode around 90–100%, suggesting: "Most patients are borderline-hypoxemic or otherwise hypoxic." Figure 1 empirically shows that none of the distributions are clean normal; they exhibit skewness, truncation, multimodality, and discretization. Therefore, the models RF,

ANN, KNN, QRF, and conformal prediction may also be applicable without classical linearity assumptions. The Bayesian probability model (BPM) referred to in this article is not a Bayesian Neural Network, a Gaussian Process Classifier, an NGBoost, or a Fully Hierarchical Bayesian model. It is like doing Bayesian Logistic Regression (BLR) with posterior simulation as the probabilistic base. It is this model's strength that it is interpretable, stable, and easy to justify. This BPM serves as a parametric probabilistic baseline against which distributions, free or ML-based, are compared. The working methodology flowchart is given in Figure 2.

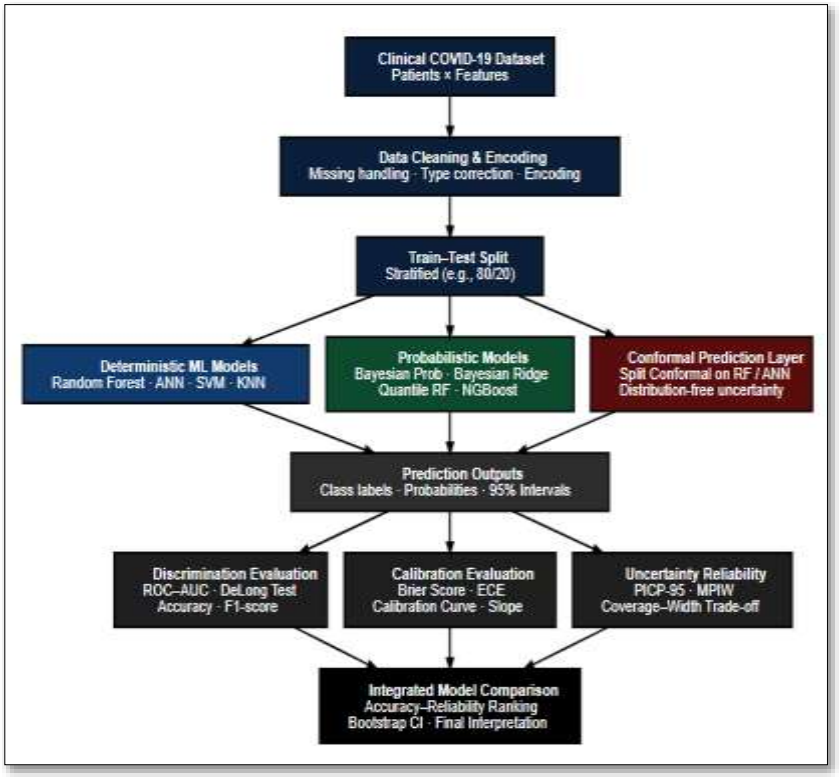


Fig. 2. Methodology Flow Chart

3.2 Random Forest, KNN, ANN, Bayesian Ridge Regression, and Quantile Regression Forest

For the Random Forest (RF) is computed using the average importance of$x$ = $\frac{1}{n_T}\sum_{i=1}^{n_T}\sum_{node\ j\in T_i\ |\ Split\ variable=x}[p_{T_i}(j)\Delta i_{T_i}(j)]$, where n_T denotes the number of trees in the forest,T_i is the ith tree, $p_{T_i}(j)=\frac{n_j}{n}$ is the fraction of sample nodes reaching the node j , $\Delta i_{T_i}(j)$ is the change of impurity (Gini) in tree i at the node j. Learning nonlinear mapping using layers of weighted sums, along with activation using the Artificial Neural Network (ANN) for binary classification (MLP) is given as below: The hidden layer is given by$z^{(1)}=\sigma(W^{(1)}x+b^{(1)})$. Output (sigmoid for binary) is given by $\hat{p}(y=1\mid x)=\sigma(W^{(2)}z^{(1)}+b^{(2)})$, where $\sigma(u)=\frac{1}{1+e^{-u}}$. The KNN distance is measured using the Euclidean distance $d(x,y)=\sum_{i=1}^{n}(y_i-x_i)^2$, where x_i and y_i are the nodes. The BRR is denoted by $P(\,y\mid\theta\,)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(y-Xw)^{\top}(y-Xw)\right]$, where y is the predictor with respect to θ, the predictor variable, σ the standard deviation, Xw is the target to the error , and T denote the transpose of the matrix. The conditional distribution function of the RF Y , given $X=x$ is given by $F(y|X=x)=P(Y\le y|X=x)=E\left(1_{Y\le y}\middle|X=x\right)$. These analogies with the RF approximation of the conditional mean $E(Y\,|X=$

x). Define an approximation to $E(1_{Y\leq y}|X=x)$by the weighted mean over the observations of $1_{Y\leq y}$, $\hat{F}(y|X=x)=\sum_{i=1}^{n} w_i(x)1_{Yi\leq y}$, using the same weights $w_i(x)$ as for RF. This approximation is at the Quantile Regression Forests (QRF).

3.3 Prediction Interval Coverage Probability (PICP 95) and Brier Score

The Prediction Interval Coverage Probability (PICP) measures how often the true value falls inside the 95% prediction interval, with a lead value of 0.956 and an acceptable prediction of 0.85–0.95. The formula for the PICP calculation is given as $\text{PICP}_{95}=\frac{1}{n}\sum_{i=1}^{n}\text{I}\left(y_i\in[L_i^{0.025},U_i^{0.975}]\right)$. The Brier Score, or simply Brier, evaluates the quality of the predicted probabilities, not just the prediction errors. If two models have the same error metrics, such as RMSE, it is challenging to decide which is better. Brier scores address this gap. A Brier score of 0 (zero) indicates higher accuracy. It captures calibration, which explains whether the probabilities are numerically meaningful; sharpness, which explains whether the probabilities are confident and correct; and reliability, which explains whether the mean of 70% is true. The Brier Score is defined as $Brier=\frac{1}{n}\sum\ (p_i-y_i)^2$, where pi is the predicted probability and $y_i\in\{0,1\}$. This score is the actual class discrimination, which differs from the MSE score. The Calibration Slope and Model Perfectness Interpretation is considered with $Calibration\ slope=\begin{cases}1 & Model\ is\ Perfect\\ <1 & Overestimanted\\ >1 & Underestimated\end{cases}$. Moreover, even if the calibration intercept is 0, the model is also deemed ideal. The various metrics, formulas, ideal values, and interpretations are given in Table 1.

Table I : METRIC DEFINITION

Metric	Formula/ Method of Calculation	Ideal value	Interpretation
AUC	Area under ROC curve	1.0	Discrimination ability
Brier Score	Brier = $(1/n)\Sigma(p_i-y_i)^2$	0.0	Probability accuracy + calibration
PICP-95	PICP = $(1/n)\Sigma 1_{(y_i\in[L_i,U_i])}$	$\approx$ 0.95	Coverage reliability of 95% intervals
MPIW	MPIW = $(1/n)\Sigma(U_i-L_i)$	As small as possible	Sharpness of intervals
Calibration Slope	Logistic: $y\sim\text{logit}(p)$ slope	$\approx$ 1	Over/under-confidence (slope)
Calibration Intercept	Logistic: $y\sim\text{logit}(p)$ intercept	$\approx$ 0	Systematic probability bias
ECE	ECE = $\Sigma_k\left(\frac{n_k}{n}\right)\lvert acc_k-conf_k\rvert$	0	Average calibration gap

4. Results and Analysis

4.1Confirmal Matrix Metrics

Table 2 shows that, in terms of accuracy, RF leads (0.8750), followed by ANN (0.8654) and SVM (0.7885); however, all the considered models perform above 0.7. The Cohen's Kappa measure, which assesses agreement between the predicted and actual classes, places ANN (0.6987) at the top, followed by RF (0.6850). Sensitivity, the proportion of actual positive cases correctly identified, places RF at the top with a value of 1, followed by SVM

(0.9000). The confusion-matrix-based statistics to assess model performance is presented in Table 2.

Table II. Confusion Matrix Statistics

Model	Accuracy	Kappa	Sensitivity	Specificity	Precision	Recall	F1
RF	0.8750	0.6850	1.0000	0.6176	0.8434	1.0000	0.9150
ANN	0.8654	0.6987	0.8857	0.8235	0.9118	0.8857	0.8986
SVM	0.7885	0.4884	0.9000	0.5588	0.8077	0.9000	0.8514
KNN	0.7692	0.4419	0.8857	0.5294	0.7949	0.8857	0.8378
BRR	0.7404	0.3670	0.8714	0.4706	0.7722	0.8714	0.8188
BPM	0.7115	0.3134	0.8286	0.4706	0.7632	0.8286	0.7945

The third place is shared by ANN (0.8857) and KNN (0.8857). Specificity, the proportion of actual negative cases identified, shows that ANN (0.8235) leads RF (0.6176) and the other models. In all measures, the Bayesian-oriented models underperform, except for specificity and kappa, which remain above 0.7.

Table III. AUC AND CI

Model	AUC	AUC_L95	AUC_U95
RF	0.92	0.85	0.98
KNN	0.79	0.70	0.88
ANN	0.88	0.81	0.96
SVM	0.85	0.76	0.93
BPM	0.79	0.70	0.88
BRR	0.79	0.70	0.88

Table III presents the AUC values and the CI, with lower (AUC_L95) and upper (AUC_U95) bounds. The AUC values show that RF is leading (0.92) with CI (0.85−0.98), followed by all the others, with ANN (0.88). Even though KNN, BPM, and Ridge are sharing the last part of the table, they are keeping a significant score (0.79) with CI (0.70−0.88). **Figure 3** shows the ROC curves of test data for the models considered.

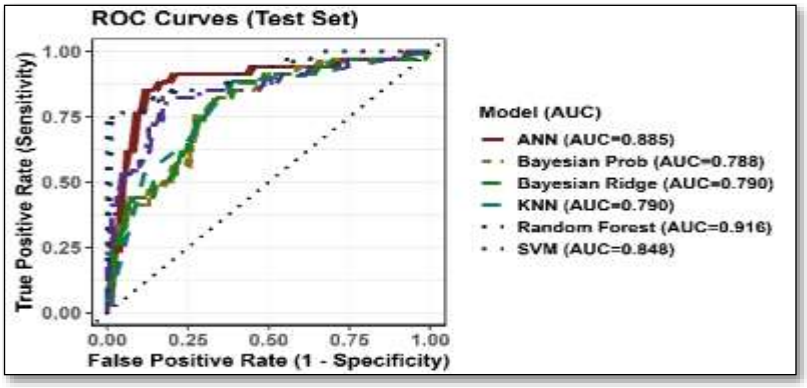


Fig. 3 ROC Curves

In Figure 3, the ROC curves for all the models, with the x-axis representing the false-positive rate and the y-axis, the true-positive rate, was above the diagonal. However, the RF model leads in decimating power, with an AUC of 0.916, followed by ANN (0.885) and SVM (0.848). The Bayesian and KNN models show moderate discrimination performance. Figure 4 shows the calibration curves with binned points and the smoothed lines of the considered curves.

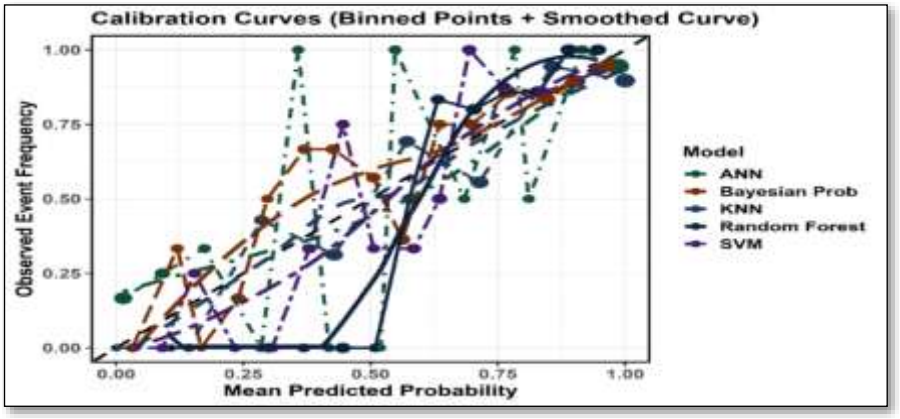


Fig. 4 Calibration Curves

In Figure 4, the calibration curves were produced with the mean predicted probability on the x-axis against the observed event frequency on the y-axis. The 45-degree diagonal line, called the ideal line, shows perfect calibration. That predicted the probability equals to observed frequency. The RF shows a good calibration at higher probability. ANN and SVM exhibit moderate oscillations but are unstable in their sensitivity and probability estimates. Despite having lower discriminative power, Bayesian methods handle uncertainty better on unseen data. Figure 5 shows the Brier Score diagram of the models considered.

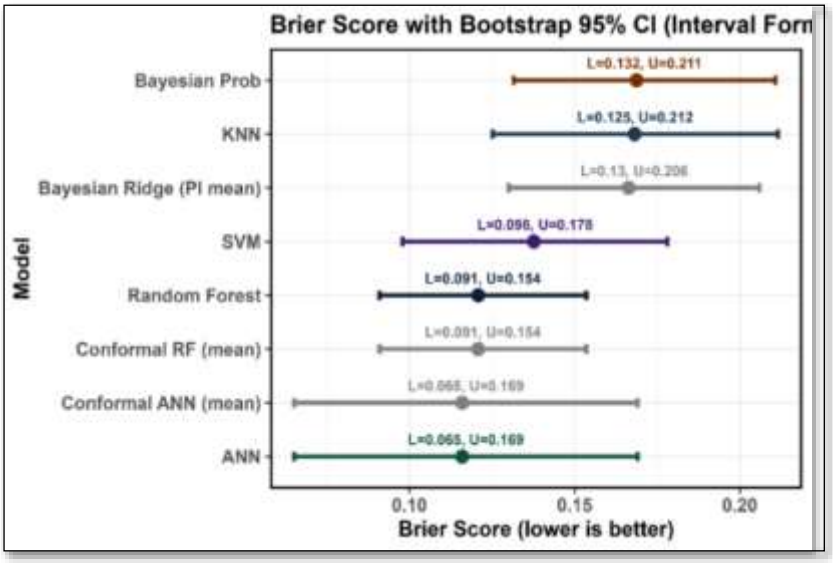


Fig. 5. Brier Score

In Figure 5, the Brier score assesses overall probability error, with lower values indicating better performance. The horizontal lines show 95% of the bootstrap, with L for lower and U for upper limits of the CI. The lower Brier scores are obtained by the ANN and Conformal ANN (0.065−0.169), indicating better probabilistic accuracy, but with wider CIs. This means that the predictions are good on average. But these predictions are unstable, especially with random data. RF and Conformal RF show a moderate Brier score (0.091) with a narrower CI (0.091−0.154), indicating a stronger balance between calibration and robustness. The BRR, the probabilities Brier score, and the CI are, respectively 0.13, (0.13−0.206), and (0.132−0.211), showing the higher error, wider CI, and struggle with nonlinear patterns. With such pure probabilistic models, the predictors are insufficient to achieve more accurate predictions. Figure 6 shows the Accuracy versus the reliability ranking graph compared with AUC and Brier score for the considered models.

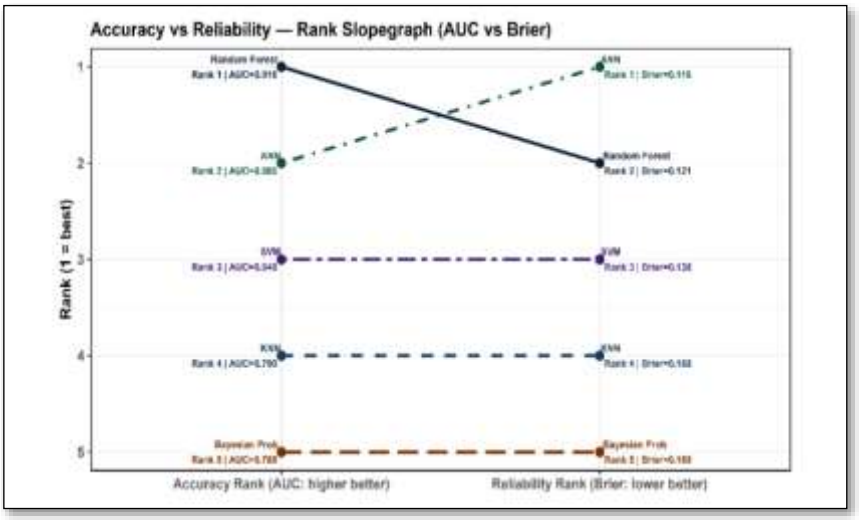


Fig. 6. Accuracy vs Reliability Ranking Slope graph (AUC vs Brier)

Figure 6 shows that RF and ANN are alternating their first two ranking positions in terms of accuracy and reliability. On the other hand, SVM, KNN, and Bayesian methods retained their positions at 3, 4, and 5, respectively.

4.1.1 AUC DeLong Pairwise Test

TABLE IV. AUC DELONG PAIRWISE TEST TABLE

Model_1	Model 2	AUC_1	AUC_2	AUC_Diff	DeLong_Z	P_Value	N_Paired	P_Adj_Holm	P_Adj_FDR	Sig_0_05
RF	BPM	0.92	0.79	0.13	3.08	0.002	104	0.012	0.004	Yes
BPM	Conformal RF	0.79	0.92	-0.13	-3.08	0.002	104	0.012	0.004	Yes
RF	BRR	0.92	0.79	0.13	2.99	0.002	104	0.012	0.004	Yes

BRR	Conformal RF	0.79	0.92	-0.13	-2.99	0.002	104	0.012	0.004	Yes
BPM	BRR	0.79	0.79	0	-0.19	0.852	104	1	1	No
RF	Conformal RF	0.92	0.92	0	0	1	104	1	1	No

Table 4 shows the AUC DeLong pairwise test results, comparing the considered models, including AUC, AUC difference, DeLong Value, and P_Value, with significance results. Mean and PI mean is considered as the point estimate for Conformal RF and BRR respectively in Table 4 AUC DeLong Pairwise Test. Table 4 shows that both Bayesian models, BPM and BRR, are significantly inferior to conformal RF, with identical effect sizes (AUC difference = -0.13). RF showing its superiority on both Bayesian models (AUC = 0.13, p = 0.002, Holm-Adjusted p = 0.013). The z value of 3 indicates a significant discriminative ability, not just marginal improvement. Compared with the Bayesian models BRR and BPM, there is no significant difference (p = 0.85). This shows that changing the Bayesian method doesn't improve discriminative capability. The crucial result is that there is no statistically significant difference between RF and conformal RF (AUC difference = 0, p = 1.0). Conformal prediction does not reduce discriminative performance. Conformal RF preserves the full classification power of RF even when adding uncertainty quantification.

4.2Error Comparison and Cross Validation

Table 5 presents the error metrics for the considered models on the Test data, including RMSE, MAE, MSE, and MedAE.

TABLE V. ERROR METRICS OF MODELS (TEST)

Model	RMSE	MAE	MSE	MedAE
RF	0.3476	0.2878	0.1208	0.2878
KNN	0.4101	0.3091	0.1682	0.3091
ANN	0.3406	0.1743	0.1160	0.1743
SVM	0.3710	0.2846	0.1377	0.2846
BPM	0.4108	0.3244	0.1688	0.3244
BRR (PI mean)	0.4078	0.3311	0.1663	0.3311
Conformal RF (mean)	0.3476	0.2878	0.1208	0.2878
Conformal ANN (mean)	0.3406	0.1743	0.1160	0.1743

Table 5, Error Metrics of the Models on Test data, shows that ANN has the lowest RMSE (0.3406), followed by RF

(0.3476). The BPM shows higher RMSE (0.4108, 0.4078) and retains its last position, preceded by KNN (0.4101). On the other hand, the MedAE reveals ANN (0.1443) as the best-performing model, followed by SVM (0.2846) and RF (0.2878). The ANN's MSE (0.1160) is lower than that of the other models, indicating the best performance among the models considered. RF (0.1208) is holding second place after ANN. Cross-validation (CV) is employed especially for ANN and KNN models. It is due to the strong sensitivity and the hyperparameter selection. RF already incorporates internal validation through out-of-bag estimation. The probabilistic and conformal models show uncertainty calibration. The ROC of KNN (CV=0.8477) with an optimal k=7, against the ANN(CV=0.8952) with an absent k value. KNN indicates good but moderate discriminative ability. This suggests that KNN captures some structure of the data, but the performance is limited. On the other hand, ANN(CV) shows superior, stable, and discriminative performance. The absence of k values reflect that ANN tuning is based on the network rather than a single value.

4.3Limitation

This work was conducted at the Sara Multispecialty Hospital in Maharashtra, India. While the models demonstrate strong internal validity, the external validity across different clinical settings may vary. Conformal prediction methods assume exchangeability between the train and test data. In a real-world scenario, when such assumptions are violated, the prediction may differ. The Bayesian models in this work rely on the prior structures and the parametric assumptions. If real-world complex data are not captured correctly, the model's accuracy may vary.

5. Conclusion and Future Work

This result shows that high predictive accuracy alone is not sufficient to determine ML model performance. In this work, while RF have highest discriminative performance, (AUC=0.92) followed by ANN (AUC =0.89). It does not guarantee reliable probability estimates. Especially, KNN and Bayesian probabilistic models BPM, and BRR show inferior reliability, evidenced by higher Brier scores (0.13−0.21). It also has wide bootstrap confidence intervals. However, conformal RF and ANN maintain competitive discrimination while exhibiting statistically valid performance. It has nominal convergence and a controlled interval, but although it achieved the lowest mean Bryer score (0.065), the ANN's CI is wide. This shows the ANN's sensitivity to data perturbations, revealing a critical reliability weakness. On the other hand, RF, particularly when augmented with conformal prediction of most balanced accuracy, calibration, stability, and uncertainty control. The current work argues that model selection must move beyond single-metric dominance to assess the reliability of word-integrated accuracy. This ensures the accuracy of the predictions that directly influence clinical decision-making and patient safety. Future research will extend this framework to temporal forecasting, multi-horizon risk prediction, and external validation across heterogeneous data.

Acknowledgment

I would like to express my sincere gratitude to Sara Multispecialty Hospital in Maharashtra, India, for providing the Data for this analysis.

References

1. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the 22nd International Conference on Machine Learning - ICML '05, 625–632. https://doi.org/10.1145/1102351.1102430
2. Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17(1), 230. https://doi.org/10.1186/s12916-019-1466-7.
3. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis (0

ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018
4. Madhan, N., Dheva Rajan, S., & Jain, M. (2025). Directing Natural Language Processing Text Similarity Challenges in Social Media with AI Techniques. In V. Goar, M. Kuri, R. Kumar, & T. Senjyu (Eds.), Advances in Information Communication Technology and Computing (Vol. 1075, pp. 425–442). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-6106-7_26
5. Madhan, N. (2024). Analysis of a Novel Hybrid Fine Tuned Weighted Harmonic Mean for Efficient Plagiarism Detection using Particle Swarm. Nanotechnology Perceptions, 20(S2). https://doi.org/10.62441/nano-ntp.v20iS2.38
6. Angelopoulos, A. N., & Bates, S. (2023). Conformal Prediction: A Gentle Introduction. Foundations and Trends in Machine Learning, 16(4), 494–591. https://doi.org/10.1561/2200000101
7. iVovk, V., Alexander, G., & Shafer, S. (2005). Algorithmic Learning in a Random World. Springer-Verlag. https://doi.org/10.1007/b106715
8. Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association, 102(477), 359–378. https://doi.org/10.1198/016214506000001437
9. Khodabakhshian, A., Puolitaival, T., & Kestle, L. (2023). Deterministic and Probabilistic Risk Management Approaches in Construction Projects: A Systematic Literature Review and Comparative Analysis. Buildings, 13(5), 1312. https://doi.org/10.3390/buildings13051312
10. Xiong, R., Shi, Y., Jing, H., Liang, W., Nakahira, Y., & Tang, P. (2024). Calibrating subjective data biases and model predictive uncertainties in machine learning-based thermal perception predictions. Building and Environment, 247, 111053. https://doi.org/10.1016/j.buildenv.2023.111053
11. Huang, Y., Chouzenoux, E., Elvira, V., & Pesquet, J.-C. (2023). Efficient bayes inference in neural networks through adaptive importance sampling. Journal of the Franklin Institute, 360(16), 12125–12149. https://doi.org/10.1016/j.jfranklin.2023.08.044
12. Xiong, Y., Ma, Y., Ruan, L., Li, D., Lu, C., Huang, L., & the National Traditional Chinese Medicine Medical Team. (2022). Comparing different machine learning techniques for predicting COVID-19 severity. Infectious Diseases of Poverty, 11(1), 19. https://doi.org/10.1186/s40249-022-00946-4
13. Luo, J., Zhang, Z., Fu, Y., & Rao, F. (2021). Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. Results in Physics, 27, 104462. https://doi.org/10.1016/j.rinp.2021.104462
14. Devi, M. S., Carmel Mary Belinda, M. J., Aruna, R., Ramesh, P. S., & Sundaravadivazhagan, B. (2023). A Novel Model for IoT Blockchain Assurance-Based Compliance to COVID Quarantine. In G. R. Kanagachidambaresan, D. Bhatia, D. Kumar, & A. Mishra (Eds.), System Design for Epidemics Using Machine Learning and Deep Learning (pp. 63–81). Springer International Publishing. https://doi.org/10.1007/978-3-031-19752-9_5
15. N Uday, K., Priya, W. D., Sureshsekar, & Sundaravadivazhagan, B. (2024). A Comparative Analysis of K-Nearest Neighbors and Support Vector Machine for Human Emotion Detection Based on Voice Patterns to Improve Accuracy. 2024 IEEE 9th International Conference on Engineering Technologies and Applied Sciences (ICETAS), 1–5. https://doi.org/10.1109/ICETAS62372.2024.11120191
16. Maad, M. M., Karan, A., Ruchi, D., Kamal Kant, H., & B, B. S. (2022). Deep Learning Techniques for COVID-19 Detection Based on Chest X-ray and CT-scan Images: A Short Review and Future Perspective. Asian Journal of Applied Sciences, 10(3), 224–231.
17. Singh, V., Khanzadeh, M., Davis, V., Rush, H., Rossi, E., Shrader, J., & Lio, P. (2024). Bayesian Binary Search (arXiv:2410.01771). arXiv. https://doi.org/10.48550/arXiv.2410.01771
18. Alzraiee, A. H., & Niswonger, R. G. (2024). A probabilistic approach to training machine learning models using noisy data. Environmental Modelling & Software, 179, 106133. https://doi.org/10.1016/j.envsoft.2024.106133
19. Ajirak, M., Ravishankar, A., & Djuric, P. M. (2025). Uncertainty Quantification in Probabilistic Machine Learning Models: Theory, Methods, and Insights (arXiv:2509.05877). arXiv. https://doi.org/10.48550/arXiv.2509.05877

20. Wang, J. (2024). Exploring Probabilistic Models for Semi-supervised Learning (arXiv:2404.04199). arXiv. https://doi.org/10.48550/arXiv.2404.04199
21. Ghosh, R. P., & Barnett, I. (2025). Bayesian Neural Networks vs. Mixture Density Networks: Theoretical and Empirical Insights for Uncertainty-Aware Nonlinear Modeling (arXiv:2510.25001). arXiv. https://doi.org/10.48550/arXiv.2510.25001
22. Tyralis, H., & Papacharalampous, G. (2024). A review of predictive uncertainty estimation with machine learning. Artificial Intelligence Review, 57(4), 94. https://doi.org/10.1007/s10462-023-10698-8
23. Dew, R., Padilla, N., Luo, L. E., Oblander, S., Ansari, A., Boughanmi, K., Braun, M., Feinberg, F., Liu, J., Otter, T., Tian, L., Wang, Y., & Yin, M. (2024). Probabilistic Machine Learning: New Frontiers for Modeling Consumers and their Choices. International Journal of Research in Marketing, S0167811624000995. https://doi.org/10.1016/j.ijresmar.2024.11.002
24. Kapoor, M. (2024). Probabilistic Machine Learning and Artificial Intelligence. Spectrum of Emerging Sciences, 3(2), 29–36. https://doi.org/10.55878/SES2023-3-2-5
25. Nimy, E., Mosia, M., & Chibaya, C. (2023). Identifying At-Risk Students for Early Intervention—A Probabilistic Machine Learning Approach. Applied Sciences, 13(6), 3869. https://doi.org/10.3390/app13063869
26. Marvasti, A. E., Marvasti, E. E., Bagci, U., & Foroosh, H. (2021). Maximum Probability Theorem: A Framework for Probabilistic Machine Learning. IEEE Transactions on Artificial Intelligence, 2(3), 214–227. https://doi.org/10.1109/TAI.2021.3086046
27. Sprangers, O., Schelter, S., & De Rijke, M. (2021). Probabilistic Gradient Boosting Machines for Large-Scale Probabilistic Regression. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 1510–1520. https://doi.org/10.1145/3447548.3467278