2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Detecting Uncertainty of Mortality Prediction Using Confident Learning

Zahra Shakeri Hossein Abad[1], and Joon Lee[2]

Abstract— **Early mortality prediction is an actively researched problem that has led to the development of various severity scores and machine learning (ML) models for accurate and reliable detection of mortality in severely ill patients staying in intensive care units (ICUs). However, the uncertainty of such predictions due to irregular patient sampling, missing information, or high diversity of patient data has not yet been adequately addressed. In this paper, we used confident learning (CL) to incorporate sample-uncertainty information into our mortality prediction models and evaluated the performance of these models using a large dataset of 139,367 unique ICU admissions within the eICU Collaborative Research Database (eICU-CRD). The results of our study validate the importance of uncertainty quantification in patient outcome prediction and show that the state-of-the-art ML models augmented with CL are more robust against epistemic error and class imbalance.**

I. INTRODUCTION

In recent years, a large body of work has been devoted to applying machine learning (ML) and artificial intelligence (AI) to predict patient outcome, from which two lines of research stand out. First, studies that utilize ML to extract clinically relevant information from complex electronic health records (EHRs) to improve the accuracy of medical diagnostic models and assist clinicians in establishing a prognosis [1]–[5]. Second, research that argues trust in human-AI collaboration and investigates the role of human clinicians in interpreting the decisions made by medical ML models [6]–[9]. Although these two lines of research have advanced the field of patient outcome prediction considerably, quantifying and communicating the uncertainty of predictive models, mainly caused by uncertain samples, remain an open and challenging area of research. By uncertain samples, we refer to samples that are difficult to correctly classify. This difficulty is class-conditional and could stem from multiple sources such as noise, dataset shift, bias, or missing information [8] (Figure 1). The presence of challenging/uncertain cases in a dataset used to train predictive models introduces two main problems that can impact the prospective applications of these models: (1) unrobust evaluation of the learning models, and (2) the lack of reliability of decisions made upon these results.

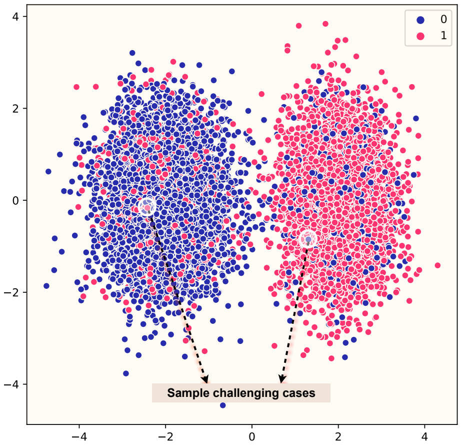


Fig. 1: A sample thematic view of challenging cases (caused by aleatoric uncertainty) in a binary classification task.

In recent years, several attempts have been made to address predictive uncertainty, including prediction intervals [10], conformal inference [11], and ensembling [12]. However, the majority of these methods are model-dependent and are subject to epistemic error and class-imbalance bias. Moreover, quantification/communication of uncertainty is not adequately addressed in mortality prediction literature, despite its critical importance in patient care and the frequency of its application in critical care decision making. In this paper, we utilize confident learning (CL) [13], a model- and data-agnostic approach to characterize the uncertainty of samples when predicting mortality for critically ill patients within 24 hours of ICU admission. CL is robust to heterogeneous and imbalanced class distributions and disambiguates model uncertainty from sample uncertainty while estimating the joint distribution of uncertain and certain labels. Our main goal in this paper is two-fold:

Pruning– Detecting challenging (i.e. uncertain) cases through multiple iterations of cross-validation and train and evaluate the machine learning models on a pruned dataset (challenging cases excluded). This objective will provide new insights into both the deficiency of existing performance metrics in incorporating uncertain samples and the importance of clean training datasets in developing predictive models.

Uncertainty prediction– Converting the binary problem of mortality prediction to a multi-class classification problem and developing ML models to detect challenging cases. This will help flag patients with highly uncertain predictions who need a second opinion.

In Section II, we first formalize the problem we are addressing and define used notations and concepts, followed by an overview of the CL approach. In Section III, we present experimental evaluation and the results. We conclude the paper by discussing the main implications of our study, with possible suggestions for future work in Section IV.

This work was supported by the Libin Cardiovascular Institute of Alberta and the Cumming School of Medicine, University of Calgary. Also, this work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2014-04743).

[1] Zahra Shakeri Hossein Abad is with Data Intelligence for Health Lab, Cumming School of Medicine, University of Calgary, Canada. zshakeri@ucalgary.ca

[2] Joon Lee is with Data Intelligence for Health Lab and Departments of Community Health Sciences and Cardiac Sciences, Cumming School of Medicine, University of Calgary, Canada. joonwu.lee@ucalgary.ca

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

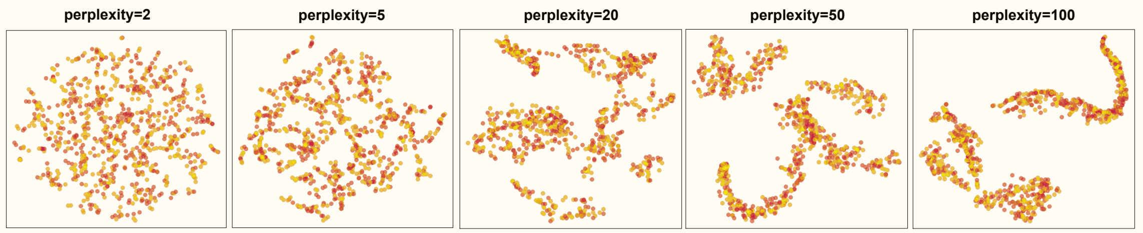


Fig. 2: t-SNE plots for different perplexities, color-coded by prediction outcomes– all correctly classified, and all misclassified. The perplexity can be interpreted as a smooth measure of the effective number of neighbors.

II. Uncertainty Detection using Confident Learning

A. Problem Formalization

In the context of binary classification with possibly uncertain labels (i.e. challenging cases), let define each sample s_i of dataset $\mathcal{D} = \{s_1, s_2, \ldots, s_n\}$ in the form of $\langle x_i, \tilde{y_i} \rangle$, where without loss of generality $x_i \in \mathbb{R}^d$ indicates the feature space of all critically ill patients (with all categorical variables are binary encoded) and $\tilde{y_i} \in \{0, 1\}$ defines the binary output. Our binary classifiers are trained to predict whether a patient will die ($\tilde{y_i} = 1$) or not ($\tilde{y_i} = 0$). The problem we address in this paper is detecting cases (i.e. s_i's) for which there might be a latent label y_i^* that, with a high probability, is different from $\tilde{y_i}$, indicating the case uncertainty. While, in the patient outcome datasets, the chance of having samples for which $\tilde{y_i} \neq y_i^*$ is very slim, we use this technique to detect prediction uncertainties and flag uncertain samples of the dataset. Thus, latent label y^* in this study refers to an indicator of sample uncertainty. After detecting uncertain samples, we train and test our ML models on two variations of $\mathcal{D}_{train}$: (1) all samples for which $\tilde{y_i} \neq y_i^*$ excluded, and (2) $\langle x_i, \tilde{y_i} \rangle$ such that $\tilde{y_i} \in \{0, 1, 3\}$, where ($\tilde{y_i} = 3$) means the sample (patient) is hard to classify.

B. Uncertainty Detection

The confident learning (CL) algorithm we utilized in this study [13] is built on the commonly used assumption that label uncertainty is class-conditional, and it can be identified based on the class labels, not the data [14]–[16]. For example, in mortality prediction, a patient with a unique clinical presentation might be more likely to be misclassified as 'expired', while the correct class is 'alive'. This approach is model-agnostic and does not associate any specific loss function with the prediction model used for detecting uncertain labels. To detect the samples for which $\tilde{y} \neq y^*$ (uncertain samples in our scenario), CL calculates the joint distribution of label uncertainty (i.e. $p(\tilde{y}, y^*)$) for every pair of $(\tilde{y}, y^*)$ in the dataset. For example, for our mortality prediction task, the algorithm estimates $p(0, 1)$, $p(1, 0)$, $p(0, 0)$, and $p(1, 1)$, where $p(i, i)$ shows the joint distribution of certain samples. To estimate the joint distribution of label uncertainty, CL disambiguate epistemic uncertainty (i.e. model errors) from aleatoric uncertainty (i.e. sample uncertainty), without any prior knowledge about the distribution of uncertain or latent labels (y^*) and finds samples that are likely to belong to y^*. Moreover, when calculating the joint probability, CL uses thresholding to mitigate the impact of higher probabilities caused by class-imbalance (e.g. larger thresholds for the majority classes). After detecting the uncertain samples, we re-trained our predictive models using a training set: (1) without uncertain samples (i.e. pruned) and (2) with uncertain samples flagged as 'challenging' (i.e. three-class classification).

III. Experimental Evaluation

A. Dataset and Data Preparation

The data for this study was obtained from the eICU Collaborative Research Database (eICU-CRD) [17], a multicenter critical care database supported by Philips Healthcare and the Laboratory for Computational Physiology [18] at the Massachusetts Institute of Technology. eICU-CRD comprises 200,859 ICU stays, from 166,355 hospital stays for 139,367 unique patients admitted to one of 335 ICUs at 208 hospitals across the United States between 2014 and 2015.

To develop the predictive mortality prediction models, we included the 32 variables that are used to calculate the Acute Physiology and Chronic Health Evaluation (APACHE) IV [19] score for estimating patient severity of illness. These variables include patient demographics, ICU admission diagnosis, chronic health condition, the elective surgery status, admission source, and physiologic and laboratory variables from the first 24 hours of the ICU stay. We used one-hot encoding to convert categorical data to dummy variables and to minimize the effect of previous ICU admissions for patients with multiple ICU stays, only the first stay was included in the analysis. To mitigate the bias resulting from nonrandom missing data (missing values rates from 0.1% to 79%), we used Multivariate Imputation by Chained Equations (MICE) [20]. The continuous features of the resulting arrays were then standardized into z-scores by subtracting the mean and scaling each feature to unit variance.

To address the class imbalance problem of the eICU dataset in terms of the distribution of in-hospital mortality outcome (death: 91% (118,994), alive: 9% (11,792)), we employed the Synthetic Minority Over-sampling TEchnique-Nominal Continuous (SMOTE-NC) [21] approach, which oversamples the minority classes by creating synthetic samples based on feature-space (rather than data-space).

B. Model Development

To mitigate the risk of over-fitting and ensure that our results are not biased towards a specific learning algorithm,

TABLE I: Comparison of the normal and pruned datasets for different predictive models.

Models	Precision		Recall		Specificity		F1		AUC		IBA*	
	Mixed	Pruned	Mixed	Pruned	Mixed	Pruned	Mixed	Pruned	Mixed	Pruned	Mixed	Pruned
Logistic Regression	89%	**95%**	85%	**93%**	58%	**73%**	87%	**94%**	82%	**95%**	49%	**69%**
KNN	90%	**95%**	76%	**90%**	71%	**84%**	81%	**92%**	79%	**92%**	54%	**76%**
Random Forest	90%	**95%**	91%	**95%**	44%	**63%**	91%	**95%**	87%	**96%**	39%	**62%**
XGBoost	91%	**96%**	90%	**96%**	54%	**65%**	90%	**96%**	87%	**96%**	41%	**64%**
AdaBoost	90%	**95%**	85%	**93%**	63%	**78%**	87%	**94%**	84%	**96%**	54%	**73%**
ExtraTrees	89%	**95%**	90%	**95%**	41%	**54%**	90%	**95%**	85%	**96%**	36%	**53%**

*IBA: Index of Balanced Accuracy

we developed and evaluated a representative set of standard ML classifiers, including generalized linear (Logistic Regression (LR)), kernel-based (Support Vector Machines (SVM)), decision-tree based (Random Forest (RF), AdaBoost, XGBoost, and ExtraTrees), and sample-based (K-Nearest Neighbours (KNN)) classifiers. Hyperparameters for each method were determined using 10-fold cross-validation Bayesian Optimization [22]. For each learning pipeline, we first split the data into training and test subsets. We then applied each of the imputation, standardization, and oversampling processes to the corresponding training set.

C. Uncertainty Detection and Evaluation Results

1) Preliminary Analysis: As mentioned earlier, in this study, by uncertain samples, we refer to samples that are more challenging to classify and may need to be flagged for review by domain experts. Thus, before applying CL to our learning phase, we first trained and tested all the models listed in Section III-B and, for each sample, identified the proportion of models that misclassified the sample (i.e. difficulty index [23]). Out of 26,158 samples in our testing set, 3.3% (866) of patients were misclassified by all the classifiers listed in Table I, from which 439 samples belong to $y_i^* = 0$ and 427 belong to $y_i^* = 1$. For the paired comparisons, this number changes from 6% (1,587, ET-KNN) to 10% (2,522, LR-AdaBoost), with the majority of the misclassified samples are from the 'expired' category. This indicates the difference in the discrimination ability of learning models to handle challenging cases. To further investigate this, we analyzed the testing set for concept drift [24] and applied the t-distribution Stochastic Neighbour Embedding (t-SNE) algorithm for different values of perplexity [25] to reveal the structure of the datasets comprised of samples that were misclassified or correctly classified by all of the classifiers. Our results show no covariate shift and, as illustrated in Figure 2, there is no distribution change in the testing set.

2) Prediction Results: Table I presents the comparisons between the performance of the classifiers for two variants of the training dataset (i.e. mixed and pruned). As shown, classifiers trained on the pruned dataset consistently performed better compared to classifiers trained on the mixed dataset across all classifiers, with over 4% improvement in precision, recall, F1, and AUC metrics. Moreover, the IBA score improvement implies a more balanced contribution of both classes into the overall model's accuracy. These experimental results substantiate the value of cleaned data (pruned data) when predicting patient outcomes using ML.

We also defined a third class called 'challenging' that represents all the uncertain samples detected by CL. In Table II, we report the results of our best performing classifier in detecting the uncertain cases and compare the performance of the model in predicting mortality in both binary and multi-class classification tasks. In comparison, to predict mortality, the model trained on the dataset with flagged 'uncertain cases' outperforms the binary classifier, with 4%, 15%, and 8% improvement in precision, recall, and F1 score of the positive class, respectively. Adding the knowledge of label uncertainty to the classifiers resulted in a significantly lower number of false negatives (FNs) and a slightly lower number of false positives (FPs), indicating the need for detecting uncertain cases before training predictive models for mortality prediction (figures 3a and 3b).

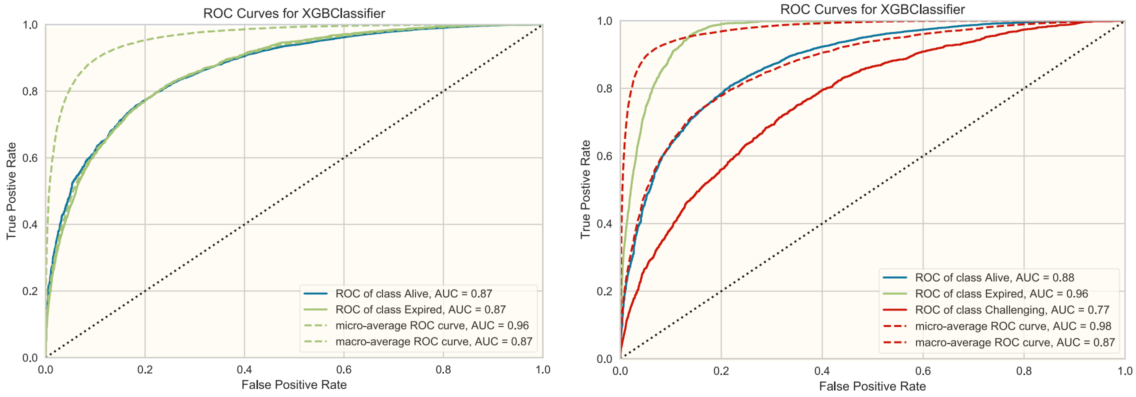
IV. Discussion and Conclusions

In this paper, we applied confident learning to the mortality prediction problem and evaluated the performance of five machine learning models (with different architectures) on two training datasets (i.e. mixed and uncertain samples pruned). The results of our study show that filtering out a subset of the training set with uncertain samples and training machine learning models on a clean dataset consistently improved the performance of our models in predicting mortality for critically ill patients. The significance of these results lies not only in improving the accuracy of models but in increasing the confidence, quality, and interpretability of clinical decisions made based on these results. Moreover, we incorporated the information of sample uncertainty into the training phase by defining a third class called 'challenging' and evaluated the performance of the XGBoost model in predicting mortality and challenging samples. As the samples assigned to the 'challenging' category were flagged based on the joint probability of aleatoric and epistemic uncertainties, the results of this multi-class classification task not only improved the discrimination ability of our models but can help identify patients for whom more information is required for better planning and clinical decision making.

These findings motivate the need for evaluation metrics that incorporate sample uncertainty into their performance quantification. The widely used metrics such as the area under the receiver operating characteristic curve (AUC), precision, recall, specificity, and F1, despite their simplicity, weight all samples equally and do not reflect the ability of predictive models in classifying uncertain samples. Also, the prospective evaluation of using confident learning in quantifying and communicating uncertainty in medical ML models requires further investigation, as a significant investment may be needed to integrate and deploy the model into real-world contexts considering the issues related to

TABLE II: Comparison of the mixed and flagged datasets for XGBoost classifier.

Models	**Precision**		**Recall**		**Specificity**		**F1**		**AUC**		**IBA**	
	Mixed	Flagged	Mixed	Flagged	Mixed	Flagged	Mixed	Flagged	Mixed	Flagged	Mixed	Flagged
Alive	95%	94%	94%	95%	50%	49%	94%	95%	87%	88%	49%	49%
Expired	46%	**50%**	50%	**65%**	94%	**96%**	48%	**56%**	87%	**96%**	45%	**60%**
Challenging	–	31%	–	14%	–	98%	–	20%	–	77%	–	13%
Micro Average	91%	89%	90%	90%	54%	54%	90%	89%	96%	98%	41%	48%



(a) Two-class classification AUC (mixed dataset)

(b) Three-class classification AUC

Fig. 3: The performance of mortality prediction for mixed and labeled challenging cases

data quantity and quality in different clinical workflows [26], [27]. Finally, as the uncertain cases in this study were identified and justified based only on theoretical methods (i.e. augmented ML models), the validation of these samples by domain experts (e.g. physicians) will shed more light on the prospective application of this approach in mortality prediction of critically ill patients.

ACKNOWLEDGMENT

We thank the Philips eICU Research Institute and the MIT Laboratory for Computational Physiology for providing the eICU-CRD used in this study.

REFERENCES

[1] J. Ramon, D. Fierens, F. Güiza, G. Meyfroidt, H. Blockeel, M. Bruynooghe, and G. Van Den Berghe, “Mining data from intensive care patients,” *Advanced Engineering Informatics*, vol. 21, no. 3, pp. 243–256, 2007.

[2] Z. Obermeyer and E. J. Emanuel, “Predicting the future—big data, machine learning, and clinical medicine,” *The New England journal of medicine*, vol. 375, no. 13, p. 1216, 2016.

[3] R. A. Taylor, J. R. Pare, A. K. Venkatesh, H. Mowafi, E. R. Melnick, W. Fleischman, and M. K. Hall, “Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data–driven, machine learning approach,” *Academic emergency medicine*, vol. 23, no. 3, pp. 269–278, 2016.

[4] Z. S. H. Abad, D. Maslove, and J. Lee, “Predicting discharge destination of critically ill patients using machine learning,” *IEEE Journal of Biomedical and Health Informatics*, 2020.

[5] C.-Y. Hung, W.-C. Chen, P.-T. Lai, C.-H. Lin, and C.-C. Lee, “Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database,” in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2017, pp. 3110–3113.

[6] O. Asan, A. E. Bayrak, and A. Choudhury, “Artificial intelligence and human trust in healthcare: Focus on clinicians,” *Journal of medical Internet research*, vol. 22, no. 6, e15154, 2020.

[7] J. Lee, “Is artificial intelligence better than human clinicians in predicting patient outcomes?” *Journal of Medical Internet Research*, vol. 22, no. 8, e19918, 2020.

[8] B. Kompa, J. Snoek, and A. L. Beam, “Second opinion needed: Communicating uncertainty in medical machine learning,” *NPJ Digital Medicine*, vol. 4, no. 1, pp. 1–6, 2021.

[9] S. Gaube, H. Suresh, M. Raue, A. Merritt, S. J. Berkowitz, E. Lermer, J. F. Coughlin, J. V. Guttag, E. Colak, and M. Ghassemi, “Do as ai say: Susceptibility in deployment of clinical decision-aids,” *npj Digital Medicine*, vol. 4, no. 1, pp. 1–8, 2021.

[10] H. Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of statistical planning and inference*, vol. 90, no. 2, pp. 227–244, 2000.

[11] G. Shafer and V. Vovk, “A tutorial on conformal prediction.,” *Journal of Machine Learning Research*, vol. 9, no. 3, 2008.

[12] Z. Toth, Y. Zhu, and T. Marchok, “The use of ensembles to identify forecasts with small and large uncertainty,” *Weather and Forecasting*, vol. 16, no. 4, pp. 463–477, 2001.

[13] C. G. Northcutt, L. Jiang, and I. L. Chuang, “Confident learning: Estimating uncertainty in dataset labels,” *arXiv preprint arXiv:1911.00068*, 2019.

[14] J. Goldberger and E. Ben-Reuven, “Training deep neural-networks using a noise adaptation layer,” 2016.

[15] D. Angluin and P. Laird, “Learning from noisy examples,” *Machine Learning*, vol. 2, no. 4, pp. 343–370, 1988.

[16] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari, “Learning with noisy labels.,” in *NIPS*, vol. 26, 2013, pp. 1196–1204.

[17] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, “The eicu collaborative research database, a freely available multicenter database for critical care research,” *Scientific data*, vol. 5, 2018.

[18] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, “Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, no. 23, e215–e220, 2000.

[19] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, “Acute physiology and chronic health evaluation (apache) iv: Hospital mortality assessment for today?s critically ill patients,” *Critical care medicine*, vol. 34, no. 5, pp. 1297–1310, 2006.

[20] S. v. Buuren and K. Groothuis-Oudshoorn, “Mice: Multivariate imputation by chained equations in r,” *Journal of statistical software*, pp. 1–68, 2010.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[22] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms,” in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[23] Z. S. H. Abad, A. Kline, and J. Lee, “Evaluation of machine learning-based patient outcome prediction using patient-specific difficulty and discrimination indices,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2020, pp. 5446–5449.

[24] H. Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of statistical planning and inference*, vol. 90, no. 2, pp. 227–244, 2000.

[25] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[26] Z. Obermeyer and E. J. Emanuel, “Predicting the future—big data, machine learning, and clinical medicine,” *The New England journal of medicine*, vol. 375, no. 13, p. 1216, 2016.

[27] M. Motwani, D. Dey, D. S. Berman, G. Germano, S. Achenbach, M. H. Al-Mallah, D. Andreini, M. J. Budoff, F. Cademartiri, T. Q. Callister, *et al.*, “Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: A 5-year multicentre prospective registry analysis,” *European heart journal*, vol. 38, no. 7, pp. 500–507, 2017.