Peer-Review

Munteanu, Briana. 2025. “Machine Learning for Assessing the Role of Immunoglobulin Free Light Chains in the Identification of Circulatory Diseases and Survival Prediction.” *Journal of High School Science* 9 (4): 373–93. https://doi.org/10.64336/001c.154326

NOTE: The comments and discussion is not in chronological order. This is because some reviews and discussions took place outside the platform.

1.I am not convinced. See for example, https://doi.org/10.3324/haematol.2024.285531, where the free light chain cut off can distinguish between time to progression to multiple myeloma. There are probably other manuscripts in the public domain that can do this with other conditions such as HF. Even if there are not, the fact remains that a binary classification of this biomarker can predict TPP of related disease. Given this, why is an ML algorithm needed at all, especially when (broadly) the algorithm has a 25% false negative rate and a 20% false positive rate? How does this improve on current cut-off diagnostics? How is it better to predict with (ML) mediocre certainty, the potential for progression, versus assigning a binary risk to such a progression? It does not make any different to the pharmacotherapy; since the regimen in the event of elevated risk (or a range of TTP) is the same regardless of whether the risk elevates to 110 or 120 in terms of FLC level. In other words, even if the ML algorithm predicts the TPP within 10% accuracy, whereas the FLC predicts it to within 20%, there is no change in the treatment. Please discuss and explain in detail in the manuscript with adequate references such as the one cited in this comment and the ones below.
2.For heart disease, “…….those in the top quartile having an unadjusted risk of mortality more than twice that of those in the lowest quartile (hazard ratio: 2.38; $p < 0.0001$)…….” Again, there seems to be enough evidence to assign FLC levels to risk. Goes back to point 1, see reference below. Discuss in the manuscript.
3.A limitation of using FLC as a ‘settled’ biomarker is that one may then miss or under-research associated markers that are easier to determine and may be detectable earlier in the disease progression event (see: https://doi.org/10.1038/s41408-025-01340-7) Discuss in the manuscript.
4.Discuss whether the features that were chosen for your ML model exhibit multi-collinearity. Present a multicollinearity heat map and VIF values. If features with high VIF were used nevertheless, discuss why, and why dimension reduction techniques such as PCA were not used in that event. Discuss implications on the metrics.
5.I find figure 5 hard to believe. Are you saying that every (approx) 100 days, you generated multiple confusion matrices by varying the classification threshold and calculated AUC on that day for circulation and non-circulation events? Could you provide a detailed explanation in the manuscript as to how this was done in your ML algorithm?
6.AUC is a threshold independent metric to compare multiple models, not compare multiple temporally separated events in the same model. If the AUC is greater for ID of circulatory deaths than deaths due to other causes (DDTOC), then the model is more accurate in ID circulatory deaths than (DDTOC). Hence, what you are essentially saying in Figure 5 is that the (relative) inaccuracy of the model to ID DDTOC can be used to differentiate between circulatory deaths and DDTOC. I do not think this is a valid scientific proposition. Please discuss in the manuscript.
7.What happens if the test AUC falls exactly between the circulatory and the DDTOC on any given day? What is that death associated with?
8.Lastly I did not understand what “censored” meant (i may have missed it if you described it in the manuscript), hence I did not understand Figures 1 and 4. Please describe in more detail in the manuscript.
https://doi.org/10.1182/blood-2007-08-108357
https://doi.org/10.1038/s41408-025-01289-7

https://doi.org/10.1016/j.jchf.2015.03.014
https://doi.org/10.2337/dc13-2227

---

1. **I am not convinced. See for example, https://doi.org/10.3324/haematol.2024.285531, where the free light chain cut off can distinguish between time to progression to multiple myeloma. There are probably other manuscripts in the public domain that can do this with other conditions such as HF. Even if there are not, the fact remains that a binary classification of this biomarker can predict TPP of related disease. *Given this, why is an ML algorithm needed at all, especially when (broadly) the algorithm has a 25% false negative rate and a 20% false positive rate? How does this improve on current cut-off diagnostics? How is it better to predict with (ML) mediocre certainty, the potential for progression, versus assigning a binary risk to such a progression?* It does not make any different to the pharmacotherapy; since the regimen in the event of elevated risk (or a range of TTP) is the same regardless of whether the risk elevates to 110 or 120 in terms of FLC level. In other words, even if the ML algorithm predicts the TPP within 10% accuracy, whereas the FLC predicts it to within 20%, there is no change in the treatment. Please discuss and explain in detail in the manuscript with adequate references such as the one cited in this comment and the ones below.**

   Answer to reviewers:

   The increasing interest in applying machine learning (ML) to biomarker-driven diagnostics has raised valid questions regarding its clinical utility, particularly when weighed against the performance and interpretability of established diagnostic cut-offs. Clinicians may be justifiably skeptical of adopting a “black box” algorithm whose predictions offer no mechanistic transparency and whose marginal gains in predictive resolution do not affect downstream therapeutic decisions (Kelly et al., 2019). However, the key strengths of ML models are their ability to model complex, non-linear, and continuous relationships without relying on hard cutoffs like traditional rule-based systems (Hastie, T., Tibshirani, R., & Friedman, J., 2009). This may help avoid situations where a patient just below a cut-off is wrongly classified as “low risk” despite having a worrisome overall profile.

   For example, the authors referenced by the reviewers (Akhlaghi, 2024) evaluated the progression rates using only two-patient groups, based on their free light chain ratio (FLCr): those above or equal to 100 and those below 100. This binary classification fails to capture meaningful differences within each group. Patients with an FLCr of 120, for example, were treated equivalently to those with an FLCr of 100. More critically, individuals with an FLCr of 99 - just below the threshold - are predicted to have substantially different progression risks than those at 100, despite being extremely close in terms of FLCr .

   Machine learning (ML) algorithms are particularly valuable in such scenarios because they do not rely on a priori data partitioning and can learn complex patterns that traditional threshold-based methods might miss. As demonstrated in Figure 4 of the manuscript, our approach enabled personalized risk estimation by leveraging the full range of available data. Moreover, the performance of the algorithm can be enhanced with the inclusion of additional clinical features, reflecting a key strength of ML: its capacity to integrate new data over time. For example, combining FLC measurements with omics data, bone marrow cytometry, imaging features, or longitudinal biomarker trends could generate composite risk profiles that go beyond simple FLCr cut-offs alone (Mateos et al., 2020). In this way, ML can move beyond merely refining existing predictions to identifying novel disease subtypes, forecasting atypical progression trajectories, and informing trial stratification - applications that go beyond simple diagnostics.

In the revised manuscript I addressed the following comment in the section **1.2 Role of Machine Learning Algorithms in Disease Progression Prediction:**

The increasing interest in applying machine learning (ML) to biomarker-driven diagnostics has raised valid questions regarding its clinical utility, particularly when weighed against the performance and interpretability of established diagnostic cutoffs. Clinicians may be justifiably skeptical of adopting a "black box" algorithm whose predictions offer no mechanistic transparency and whose marginal gains in predictive resolution do not affect downstream therapeutic decisions (Kelly et al., 2019). However, one key strength of ML models is their ability to model complex, non-linear, and continuous relationships without relying on hard cutoffs like traditional rule-based systems (Hastie, T., Tibshirani, R., & Friedman, J., 2009). This may help avoid situations where a patient just below a cutoff is wrongly classified as "low risk" despite having a worrisome overall profile. Statistical methods that rely on separating patients in groups and then determining a common prognosis model for each group are dependent on the subjectivity of groups' definition and do not provide the means for individual characterization.

and the following comment in the section **4. Conclusions:**

Machine learning (ML) algorithms are capable to capture complex, nonlinear patterns in data that traditional threshold-based methods may fail to detect. As demonstrated in Figure 4 of the manuscript, our approach enabled personalized risk estimation by leveraging the full range of available data. Moreover, the performance of the algorithm can be enhanced with the inclusion of additional clinical features, reflecting a key strength of ML: its capacity to integrate new data over time. For example, combining FLC measurements with omics data, bone marrow cytometry, imaging features, or longitudinal biomarker trends could generate composite risk profiles that go beyond simple FLCr cut-offs alone (Mateos et al., 2020). In this way, ML can move beyond merely refining existing predictions to identifying novel disease subtypes, forecasting atypical progression trajectories, and informing trial stratification - applications that go beyond simple diagnostics.

References

Akhlaghi, T., Maclachlan, K., Korde, N., Mailankody, S., Lesokhin, A. M., Hassoun, H., ... & Hultcrantz, M. (2024). Evaluating serum free light chain ratio as a biomarker in multiple myeloma. *Haematologica*, *110*(2), 493.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, *17*(1), 195.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning.

Mateos, M. V., & González-Calle, V. (2017). Smoldering multiple myeloma: Who and when to treat. *Clinical Lymphoma Myeloma and Leukemia*, *17*(11), 716-722.

2. **For heart disease, "…….those in the top quartile having an unadjusted risk of mortality more than twice that of those in the lowest quartile (hazard ratio: 2.38; p < 0.0001)……." Again, there seems to be enough evidence to assign FLC levels to risk. Goes back to point 1, see reference below. Discuss in the manuscript.**

Answer to reviewers:

While quartiles (e.g., top 25% vs. bottom 25%) could show significant FLCr associations with mortality rate, they flatten the data. Everyone in the top quartile is treated the same, and the exact FLC value - whether just above the 75th percentile or way above it – is ignored. In contrast, ML models (e.g., random forests, gradient boosting) use the actual continuous

value of FLC to learn how mortality risk increases progressively with it. Moreover, ML uses the full distribution of FLC - not just above 75% or below the 25% thresholds - improving both discrimination and calibration. Dichotomization of continuous variables (like into quartiles) throws away information. In addition, while thresholds are arbitrary and dataset-dependent (i.e., 75th percentile in one cohort might be very different in another), ML models can adapt to different value distributions during training (Hastie, T., Tibshirani, R., & Friedman, J. , 2009).

In the revised manuscript I addressed this comment in the section **1.2 Role of Machine Learning Algorithms in Disease Progression Prediction,** together with the comments related to question (1):

For example, quartile-based discretization is a common statistical practice (Maeng, 2025; Jackson,2015), but it inherently flattens the data by reducing rich, continuous variation into broad categorical bins. This simplification results in a loss of granularity, as meaningful differences within each quartile are ignored, treating all values within a range as equivalent. Such an approach can obscure subtle but important trends, weaken associations, and distort relationships between variables—especially in skewed distributions where quartiles may not represent evenly spaced or meaningful value ranges. In contrast, ML algorithms can handle continuous biomarker data directly, preserving the full variability and structure of the data without imposing arbitrary thresholds or groupings. This allows them to capture complex, nonlinear relationships and subtle patterns that might be lost through quartile-based categorization. Additionally, many ML methods are robust to skewed distributions and can learn from the data as is, making them more flexible and informative in scenarios where classical statistical approaches may oversimplify or distort the underlying signals.

References

Maeng, C.V., Rögnvaldsson, S., Einarsson Long, T. *et al.* Revised free light chain reference intervals enhance risk stratification in monoclonal gammopathy of undetermined significance and reduce overdiagnosis. *Blood Cancer J.* **15**, 80 (2025).

Jackson, C. E., Haig, C., Welsh, P., Dalzell, J. R., Tsorlalis, I. K., McConnachie, A., ... & McMurray, J. J. (2015). Combined free light chains are novel predictors of prognosis in heart failure. *JACC: Heart Failure*, *3*(8), 618-625.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning.

3. **A limitation of using FLC as a 'settled' biomarker is that one may then miss or under-research associated markers that are easier to determine and may be detectable earlier in the disease progression event (see: https://doi.org/10.1038/s41408-025-01340-7) Discuss in the manuscript.**

Answer to reviewers:

While FLC has become a well-established biomarker, relying too heavily on it as a settled or definitive indicator poses several limitations. One important concern is the risk of underexploring other potentially valuable biomarkers that may be more easily measured, cost-effective, or detectable at earlier stages of disease progression. The clinical focus on FLC can inadvertently narrow the scope of investigation, thereby delaying or overlooking the identification of alternative or complementary markers that could enhance early diagnosis, risk stratification, or treatment monitoring. This issue is particularly relevant in the evolving landscape of precision medicine, where a multi-parametric approach often yields better prognostic and diagnostic accuracy. Therefore, while FLC remains essential in current clinical workflows, it is important to avoid anchoring exclusively to it. Ongoing research should aim to explore and validate complementary biomarkers that may extend or

improve upon FLC's prognostic capabilities, particularly in early-stage or ambiguous cases. ML algorithms, like the ones discussed in the present paper, can easily incorporate additional biomarkers as they become available and even determine their importance in making the prediction.

In the revised manuscript I addressed the following comment in the section **3.4 Study Limitations:**

A second limitation is that, while FLC levels can provide important information about the future evolution of the patients, relying too heavily on this single biomarker presents the risk ignoring other potentially valuable biomarkers that may be more easily measured, cost-effective, or detectable at earlier stages of disease progression. Future research should aim to explore and validate complementary biomarkers that may extend or improve upon FLC's prognostic capabilities, particularly in early-stage or ambiguous cases. ML algorithms, like the ones discussed in the present paper, can easily incorporate additional biomarkers as they become available and even determine their importance in making the prediction.

4. **Discuss whether the features that were chosen for your ML model exhibit multi-collinearity. Present a multicollinearity heat map and VIF values. If features with high VIF were used nevertheless, discuss why, and why dimension reduction techniques such as PCA were not used in that event. Discuss implications on the metrics.**

Answer to reviewers:

Multicollinearity refers to a situation where two or more features are highly correlated. This can pose challenges in some statistical models which require feature independence (e.g., linear regression).

However, in the case of XGB, RF, and SVM, the impact of multicollinearity is considerably minimized. XGB and RF are tree-based models, and they perform feature selection during the tree-building process (Hastie, Tibshirani, & Friedman, 2009). Highly correlated features may be split on once and ignored thereafter; therefore, including redundant features doesn't degrade model performance. Moreover, heatmaps and VIF add no value for tree-based models since the algorithm handles redundancy internally. SVM contains a parameter that controls regularization and plays a critical role in balancing model complexity and error tolerance. Regularization reduces the impact of redundant or correlated features and helps with noisy and overlapping data (Xu, Caramanis, & Mannor, 2009).). Thus, regularization reduces the impact of multicollinearity without needing to detect it explicitly via VIF or heatmaps. In addition, PCA or other dimensionality reduction techniques were not used with these models. XGB, RF, and SVM showed stable cross-validation scores with no signs of high variance or overfitting because parameter optimization was performed for all models. Applying PCA just to address theoretical multicollinearity - when it's not affecting model performance - would introduce complexity without clear benefit. Even more important, retaining only a few PCA components, instead of all available features, would eliminate important information from the model. Interpretability would also suffer, as linear combinations of features would lose their clinical meaning.

In the revised manuscript I addressed the following comment in the section **2.3 Modeling:**

It is worth mentioning that one important benefit of the selected ML algorithms is their robustness to data multicollinearity. This concept refers to a situation where two or more features are highly correlated, meaning one can be linearly predicted from the others with high accuracy. This can pose challenges in some statistical models, especially those that

assume feature independence (e.g., linear regression). However, in the case of XGB, RF, and SVM, the impact of multicollinearity is considerably minimized. XGB and RF are tree-based models, and they perform feature selection during the tree-building process (Hastie, Tibshirani, & Friedman, 2009). Highly correlated features may be split on once and ignored thereafter; therefore, including redundant features doesn't degrade model performance. In turn, SVM contains a parameter that controls regularization and plays a critical role in balancing model complexity and error tolerance. Regularization reduces the impact of redundant or correlated features and helps with noisy and overlapping data (Xu, H., Caramanis, & Mannor, 2009). Thus, standard dimensionality reduction and feature orthogonalization (like Principal Component Analysis) are not needed for the success of the classification.

References

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning.

Xu, H., Caramanis, C., & Mannor, S. (2009). Robustness and regularization of support vector machines. *Journal of machine learning research*, *10*(7).

5. **I find figure 5 hard to believe. Are you saying that every (approx) 100 days, you generated multiple confusion matrices by varying the classification threshold and calculated AUC on that day for circulation and non-circulation events? Could you provide a detailed explanation in the manuscript as to how this was done in your ML algorithm?**

Answer to reviewers:

To assess the model's discriminatory ability over time, we computed the dynamic AUC (Hung and Chiang, 2010) at intervals of 150 days. The metric quantifies how well a model can distinguish subjects who would experience an event by time t from those who will not. While standard AUC measures the discrimination for binary outcome, the dynamic AUC determines the model's discrimination power up to and including the time t. More important, it has the possibility to account for the censoring prior to time t, by including the inverse probability of censoring weights (IPCW). Thus, the dynamic AUC is not simply a calculation of standard AUCs at different times, but a more complex metric capable of accounting for patients who exited the study before experiencing the event. Essentially, the calculation was performed by considering all pairs made up of those who experienced the event by the time t and those who did not, weighted by the IPCW.

In the revised manuscript I addressed this comment in the section **3.3 Time-Dependent Survival Probability Model**

References

H. Hung and C. T. Chiang, "Estimation methods for time-dependent AUC models with survival data," Canadian Journal of Statistics, vol. 38, no. 1, pp. 8–26, 2010.

6. **AUC is a threshold independent metric to compare multiple models, not compare multiple temporally separated events in the same model. If the AUC is greater for ID of circulatory deaths than deaths due to other causes (DDTOC), then the model is more accurate in ID circulatory deaths than (DDTOC). Hence, what you are essentially saying in Figure 5 is that the (relative) inaccuracy of the model to ID DDTOC can be used to differentiate between circulatory deaths and DDTOC. I do not think this is a valid scientific proposition. Please discuss in the manuscript.**

Answer to reviewers:

AUC is a measure of discrimination, not accuracy. It tells how well the model distinguishes between classes - not how often it gets them right at a specific probability threshold. A higher AUC for circulatory deaths means the model can more confidently identify them within the data - it does not imply anything about accuracy of the model, which changes with the value of the probability threshold selected to separate the classes. The accuracy is evaluated on the training data as a proportion of all correctly identified cases, and details related to it are included in the confusion matrix.

Although the AUC is a threshold-independent metric commonly used to compare the discriminative ability of different models, in survival analysis the AUC can be adapted (e.g., dynamic AUC) to evaluate how well a model predicts outcomes at different time points. The reason that the model performs better in identifying circulatory deaths than DDTOC simply indicates that the available features are better discriminators of cardiovascular negative evolution than of the DDTOC. The paper does not indicate or suggests that "(relative) inaccuracy of the model to ID DDTOC can be used to differentiate between circulatory deaths and DDTOC". Such differentiation would not have any scientific value. The decision that the patient is at risk for dying of cardiovascular or any other condition should be made based on the model estimated survival probability for the moment of time of interest, as well as other clinical observations, if available.

In the revised manuscript I addressed the following comment in the section **3.3 Time-Dependent Survival Probability Model**, together with the response for question (5):

As an overall measure of model discrimination, the AUC must not be interpreted as a substitute of model accuracy for predicting which patients will survive past a certain date. While AUC measures the overall ability of the model to rank a randomly chosen patient that died before a specific date higher than a randomly chosen patient that survived past that date, the prediction whether a patient will die or survive is made based on the survival probability for that specific patient.

7. **What happens if the test AUC falls exactly between the circulatory and the DDTOC on any given day? What is that death associated with?**

Answer to reviewers:

The dynamic AUC value is a model evaluation metric for survival models, not a classification tool. It tells how well the model performs across a dataset when different probabilities thresholds are employed. The decision to predict whether a patient would survive past a specific moment of time is made based on the calculated probability of survival at that time. In the standard scenario, if the survival probability is more than 50%, then it is predicted that the patient would be alive past that moment. This threshold probability (if different than 50%) is decided by the practitioner based on the cost of making a wrong prediction.

Specifically, the fact that the AUC value of circulatory model (0.74) is higher than the AUC value of DDTOC (0.65) at day 3000 (Figure 5) only informs the user that the former model is better at separating the patients that would live from those who would die. However, the decision of whether the patient leaves, dies of cardiovascular disease, or of something else is made based on the survival probabilities, as well as other clinical considerations, if available.

In the revised manuscript I did not address this comment, as it appeared to be a clarification rather than a specific request for revision from the reviewers.

8. **Lastly I did not understand what "censored" meant (i may have missed it if you described it in the manuscript), hence I did not understand Figures 1 and 4. Please describe in more detail in the manuscript.**

Answer to reviewers:

Censoring is a fundamental aspect of survival analysis, referring to situations where the exact time of an event (i.e., death, failure, or relapse) is not observed. The most common type and the one relevant to our case is right censoring which occurs when the event hasn't happened by the end of the study or when a subject leaves the study early.

In the revised manuscript I addressed this comment in the section **2.1 Data description**, above Figure1.

I sincerely appreciate the responses to my comments and attempts to address them. However, you have circumvented part of my concerns by resorting to the usual 'black-box' critique and the increased discrimination and calibration over (say) quartiles. You have not answered satisfactorily the following nuances:

1."…….why is an ML algorithm needed at all, especially when (broadly) the algorithm has a 25% false negative rate and a 20% false positive rate? How does this improve on current cut-off diagnostics?……" I would like you to run the data from the references I cited using your algorithm and compare your results with the actual survival in those studies.
2."…….It does not make any different to the pharmacotherapy; since the regimen in the event of elevated risk (or a range of TTP) is the same regardless of whether the risk elevates to 110 or 120 in terms of FLC level. In other words, even if the ML algorithm predicts the TPP within 10% accuracy, whereas the FLC predicts it to within 20%, there is no change in the treatment……."
3."…..there seems to be enough evidence to assign FLC levels to risk……." The risk determines the treatment (not whether the risk is calibrated to a specific value in a range).
4."…….100 days, you generated multiple confusion matrices by varying the classification threshold and calculated AUC on that day for circulation and non-circulation events?…." i did not understand the answer. Did you vary the classification thresholds for each AUC point (100 days) If so, how was this actually done?
5."…….Hence, what you are essentially saying in Figure 5 is that the (relative) inaccuracy of the model to ID DDTOC can be used to differentiate between circulatory deaths and DDTOC. I do not think this is a valid scientific proposition."… you do suggest this in the manuscript but do not provide a valid answer.

I read the paper with regard to the ML methodology as well as reading through the earlier reviews. My main concerns have to do with data leakage, multicollinearity and spurious ordinality. I list them below:

1.there is data leakage by using 'chapter' (cause of death) as a predictor when outcomes are defined from it; this makes the models non-deployable at baseline and inflates performance.
2.Although tree methods mitigate collinearity, no empirical checks (correlation matrices, VIF) were reported; SVM remains vulnerable to redundant features without proper preprocessing.
3.LabelEncoder on categorical variables for SVM imposes spurious ordinality; one-hot encoding is required.

1. I do not agree with you on point 1. In objective 2, the label 'circulatory' vs 'other' is derived directly from cause of death information (chapter), which is not available at baseline. This means that to assign labels for training, you already need to know the cause of death (i.e. the future outcome). Hence, chapter is still post-outcome information, even though it may not be input per se'. Thus, the labels themselves are defined using data that would not be available at prediction time. This is still target leakage - even though it may not be construed as feature leakage, and it still leads to optimistic, non-deployable models, because you are teaching the model to predict something that is already derived from future information.

We sincerely thank the reviwers for this thoughtful follow-up and the opportunity to clarify the methodological aspects of Objective 2 from a machine learning (ML) perspective.

We fully acknowledge the concern that using the *chapter* variable (cause of death) to define the "circulatory" vs "other" outcome categories involves post-event information. However, we respectfully submit that this does **not** constitute data leakage within the framework of **supervised ML**, because *chapter* was used solely to define the target labels, not as a predictor. From an ML perspective, the use of post-outcome information is intrinsic to *label definition* in supervised learning, not to *feature generation*. In our study, the *chapter* variable was never used as a predictor and thus never entered the model as input. Instead, it served solely to assign ground-truth labels that the model was trained to predict from pre-outcome baseline features. This process adheres to the canonical supervised learning workflow, which involves three key stages (Kubsch, Krist & Wulff, 2025):

1. **Training with labeled examples**
   The model is provided with examples that include **input features** (baseline variables) and their corresponding **outcomes (labels)**. In our study, the baseline features were all variables measured prior to the outcome, and the labels ("circulatory" vs "other") were derived from the *chapter* variable solely for defining the ground truth.
2. **Learning the relationship between inputs and outcomes**
   The ML algorithm learns patterns linking the features to the known outcomes. Mathematically, this is the estimation of a function
   $$f(X) \rightarrow y,$$
   where X represents baseline features and y the known outcome label. The label guides the learning process but is **never used as an input**.
3. **Prediction and evaluation**
   After training, the model is tested on unseen examples (the held-out test set) that include only features, not outcomes. The model's predictions are then compared against the true labels to evaluate accuracy. This procedure ensures a clean separation between inputs and outcomes and prevents data leakage.

This workflow is standard across all supervised ML applications, including medical and epidemiological studies where future outcomes (e.g., death, recurrence, disease subtype) are used to label training data. The fact that labels are derived from post-event information does not imply leakage; rather, it is inherent to how supervised learning operates. Leakage occurs only when outcome information is inadvertently incorporated into **predictor variables**, which we carefully avoided. All predictors in our study were strictly baseline variables available prior to the outcome.

In summary:

 The *chapter* variable was used exclusively to define the target labels and was never included among the model inputs.

 All predictor features were measured at baseline, ensuring that the model relied only on pre-outcome information.

 The entire ML pipeline followed the standard supervised learning protocol with strict separation between training and testing data.

To reinforce our explanation, we include a paragraph from Kotsiantis (2007), who provides a comprehensive overview of supervised machine learning algorithms:

“Supervised machine learning is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances. In other words, the goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown.”

References:

Kubsch, M., Krist, C., & Wulff, P. (2025). Automation—supervised machine learning. In Applying Machine Learning in Science Education Research: When, How, and Why? (pp. 167-210). Cham: Springer Nature Switzerland.
https://library.oapen.org/handle/20.500.12657/99864

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.
https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf

In the revised manuscript, we did not add a paragraph, since it seemed to be a clarification rather than an explicit request for revision from the reviewers.

2. Please include a sentence to the effect that 'multicollinearity was assessed and all feature VIF were less than 3.25" or something similar. That will suffice.

We included the following sentence in the section 2.3 Modeling:
“We calculated the Variance Inflation Factors (VIF) for all features, and all values were below 3.25, showing that multicollinearity was adequately assessed.”

3. Please justify briefly in the manuscript using "The variable *chapter* served as the target, and using label encoding for the target does not impose spurious ordinality in SVMs; this is standard practice in classification tasks (Bisong, 2019)." and include the reference you mention in the manuscript.

We included the following sentence in the section 2.2 Data Preprocessing:
“The variable *chapter* served as the target, and using label encoding for the target does not impose spurious ordinality in SVMs; this is standard practice in classification tasks (Bisong, 2019).”

Bisong, E. (2019). Introduction to Scikit-learn. In *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners* (pp. 215-229). Berkeley, CA: Apress.

---

1. I do not agree with you on point 1. In objective 2, the label 'circulatory' vs 'other' is derived directly from cause of death information (chapter), which is not available at baseline. This means that to assign labels for training, you already need to know the cause of death (i.e. the future outcome). Hence, chapter is still post-outcome information, even though it may not be input per se'.

Thus, the labels themselves are defined using data that would not be available at prediction time. This is still target leakage - even though it may not be construed as feature leakage, and it still leads to optimistic, non-deployable models, because you are teaching the model to predict something that is already derived from future information.

2. Please include a sentence to the effect that 'multicollinearity was assessed and all feature VIF were less than 3.25" or something similar. That will suffice.

3. Please justify briefly in the manuscript using "The variable *chapter* served as the target, and using label encoding for the target does not impose spurious ordinality in SVMs; this is standard practice in classification tasks (Bisong, 2019)." and include the reference you mention in the manuscript.

Dear Dr. Shireesh Apte,

Thank you for forwarding the reviewer's further comments on our manuscript titled " Machine Learning for Assessing the Role of Immunoglobulin Free Light Chains in Identification of Circulatory Diseases and Survival Prediction ". We write to clarify the matter and to request the editor's guidance on how to address the reviewer's concerns in the second revision.

1. The reviewer requests the following: "I would like you to run the data from the references I cited using your algorithm and compare your results with the actual survival in those studies."

https://doi.org/10.3324/haematol.2024.285531,

https://doi.org/10.1182/blood-2007-08-108357,

https://doi.org/10.1038/s41408-025-01289-7,

https://doi.org/10.1016/j.jchf.2015.03.014,

https://doi.org/10.2337/dc13-2227

Unfortunately, none of the raw patient datasets used in these publications are publicly available, due to privacy concerns. Moreover, the methodology presented in our paper allows for INDIVIDUALIZED survival curve PREDICTION, while the aforementioned studies present the OBSERVED cohort survival results - a single survival curve per each group of patients. Lumping our individual result together will nullify the benefits of the currently proposed method. We kindly ask that this point be clarified with the reviewer.

Would you prefer us to explicitly state in our revised manuscript that external-cohort re-application is outside the current scope, but is planned as future work? We are eager to revise the manuscript in accordance with journal standards and the reviewer's comments, and are grateful for your guidance to ensure we respond appropriately and constructively.

2. The reviewer stated that I "have not answered satisfactorily the following nuance": "…….It does not make any different to the pharmacotherapy; since the regimen in the event of elevated risk (or a range of TTP) is the same regardless of whether the risk elevates to 110 or 120 in terms of FLC level. In other words, even if the ML algorithm predicts the TPP within 10% accuracy, whereas the FLC predicts it to within 20%, there is no change in the treatment……."

In short, the reviewer considers the merits of the proposed methodology only based on its immediate implications for the established pharmacotherapy and medical practice. We would like to emphasize that we have never suggested that the goal of this paper was to be a guideline for changing the current medical treatments. The objective of the paper is only to present a machine-learning based methodology that allows for individual patient prognosis, as a complementary approach to classical statistical analysis. The manuscript has never made the claim that proposed individual prognosis alone should, or will lead to altered pharmacotherapy protocols. The actual clinical decision for treatment remains subject to established guidelines, clinician judgment, patient context, and other factors beyond model output.

To ensure clarity and avoid any potential misinterpretation, we propose to add a paragraph in the Discussion section along the following lines:

"It is important to emphasize that the present model development and evaluation study does not propose changes to pharmacotherapy regimens, nor does it assume that more precise risk prediction automatically warrants different treatment decisions. The model's output is designed to inform risk stratification and monitoring, rather than act as a standalone justification for therapeutic modification. Clinical treatment decisions remain grounded in established care pathways, patient comorbidities, and clinician-patient discussions."

I hope that this statement will address the reviewer's concern and clarify the manuscript's intended scope.

3. The reviewer stated that I "have not answered satisfactorily the following nuance": "…..there seems to be enough evidence to assign FLC levels to risk……." The risk determines the treatment (not whether the risk is calibrated to a specific value in a range).

This comment, as the previous one, refers to the implications of the present paper on the established medical treatment. We mention again that our manuscript focuses on the development and evaluation of a predictive algorithm (machine learning-based time-to-event model) and does not make recommendations regarding changes in pharmacotherapy or treatment regimens based on individual risk assessment.

In light of this, we would appreciate your guidance on how best to address this point.

4. "…….100 days, you generated multiple confusion matrices by varying the classification threshold and calculated AUC on that day for circulation and non-circulation events?…." i did not understand the answer. Did you vary the classification thresholds for each AUC point (100 days) If so, how was this actually done?

This comment appears to be a clarification rather than a specific request for revision from the reviewers. However, to address the reviewer's concern, we would appreciate your guidance on whether the following explanation would be acceptable for the reviewer:

In a standard (static) AUC setting, the outcome is binary (event vs. no event) and the classification threshold varies over all possible values to generate a ROC curve (true positive rate vs. false positive rate). The AUC is then the area under that curve, summarizing discrimination across all thresholds. This is equivalent to aggregating the results from many confusion matrices built at different cutoffs. However, in the dynamic AUC framework (Hung & Chiang, 2010) the event status is not fixed. For each time point t (e.g., day 100), individuals are classified as those who have experienced the event by day 100 (cases) and those who have not yet experienced the event by day 100 and are still under observation (controls). Some individuals may have been censored (e.g., lost

to follow-up) before day 100. To account for that, each pair of individuals (one case, one control) is weighted by the inverse probability of censoring (IPCW), ensuring fair comparison. The dynamic AUC is computed as the probability that the model assigns a higher predicted risk score to someone who had the event before t than to someone who did not have the event by t, after applying the IPCW weights.

While multiple confusion matrices at day 100 are not explicitly generated, the dynamic AUC calculation implicitly integrates over all possible thresholds, just as standard AUC does. Thus, the concept of "varying thresholds" is inherent in the computation, but it's not done by manually generating confusion matrices.

In addition, we would like to insert the following paragraph in the Discussion section, if the reviewers agree with it:

"The dynamic AUC calculation is performed at each moment of time by considering two groups of patients: those who suffered the event by that time ("cases") and patients who survived past that time ("controls"). The main difference from the standard AUC determination is that the dynamic AUC also takes into account the fact that some of the patients might have dropped out of the study until that moment (right censoring). As we do not know if the people who left the study would have been "cases" or "controls", the people in the two groups above are "weighted" by the inverse probability of censoring (IPCW), thus ensuring that the right censored cases are accounted for. The IPCW value for each person is determined by estimating the censoring distribution from the survival times in the training data."

5. The reviewer states that "...Hence, what you are essentially saying in Figure 5 is that the (relative) inaccuracy of the model to ID DDTOC can be used to differentiate between circulatory deaths and DDTOC. I do not think this is a valid scientific proposition.… you do suggest this in the manuscript but do not provide a valid answer."

Our intention was not to imply that the relative inaccuracy of identifying deaths due to other causes (DDTOC) can be used to differentiate between circulatory and non-circulatory deaths. We explicitly addressed this in both our response letter and the revised manuscript and clarified that: (1) AUC is a measure of discrimination, not accuracy; (2) it quantifies how well the model ranks patients by risk across possible thresholds, not how accurately it classifies outcomes at a fixed threshold; (3) a higher AUC for circulatory deaths indicates that the available features discriminate better for that outcome, not that the model is more accurate or that its "inaccuracy" for DDTOC can be used diagnostically; and (4) we did not claim or imply that the difference in AUCs has predictive or scientific value in distinguishing causes of death.

We respectfully request that the editor considers our clarification in light of this context. In addition, to resolve the matter, we respectfully request the editor's guidance if would it be acceptable to add a short paragraph in the Discussion section, explicitly stating:

"The difference in the dynamic AUCs between the circulatory deaths and deaths due to other conditions reflects the overall degrees of discriminative ability of each model for these outcomes; it does not imply that a higher discriminative ability for deaths due to circulatory condition can be used as a marker that the death is due to this condition for each patient. Specifically, given the same patient for which death probabilities are calculated through both models, the decision of which condition has a higher chance of leading to the event will be made based on the two probability

values, and the physician’s observations. A model with a higher dynamic AUC is expected to be more accurate in determining the individual event probabilities.”

We are committed to revising the manuscript accordingly and would appreciate the editor’s guidance so we can do so in a way that satisfies all parties.

Thank you for your time and consideration.

Sincerely,

Briana Munteanu

Otilia C. Barbu, PhD

shireesh apte <journalofhighschoolscience@gmail.com>
Wed, Oct 29, 7:28 PM
to Briana

Dear author,

Thank you for attempting to address the reviewer's comments.

1. The reviewer is concerned that your model with a 25% false negative rate and a similar false positive rate does not improve on the binary cut off for classification of risk as 'low' or 'high'. They are not asking for lumping your individual ML results. Can you match each of your individual results with whether those individual results correctly fall into the literature binary model ?

2 and 3. The reviewer is asking that since your model's results do not affect subsequent pharmacology; then what is the use of a 'greater' discriminatory value provided by your model? If (say) your model predicted greater risk so that a different pharmacotherapy could be applied, then that discriminatory nuance would have some value. If it does not, why is a 'complementary approach to classical statistical analysis' needed? If you can find literature that suggests that dose-titration or different therapy regimens or protocols in high risk patients improve survival; that would be a way to answer this question. If you cannot find such literature, you can put this down as a limitation of your approach.

4. Your paragraph is acceptable. However, I still would like to see a different nomenclature than "AUC". Can you confirm that Hung & Chiang, 2010, used this same nomenclature "dynamic AUC" in their work? If yes, then it is acceptable. If not, then please use the term that they used.

5. Does the AUC calculation for circulatory related deaths use different (or a subset) of features than the AUC calculation for all-cause mortality? If yes, then you have a different model for these two conditions and you will need to state that. If not, the paragraph that you have proposed is acceptable.

Please let me know if point 1 is do-able.
Thank you,
Best,
Shireesh Apte
The content below is a reproduction of the reviewer's comments.

I sincerely appreciate the responses to my comments and attempts to address them. However, you have circumvented part of my concerns by resorting to the usual 'black-box' critique and the increased discrimination and calibration over (say) quartiles. You have not answered satisfactorily the following nuances:

"…….why is an ML algorithm needed at all, especially when (broadly) the algorithm has a 25% false negative rate and a 20% false positive rate? How does this improve on current cut-off diagnostics?……" I would like you to run the data from the references I cited using your algorithm and compare your results with the actual survival in those studies.

"…….It does not make any different to the pharmacotherapy; since the regimen in the event of elevated risk (or a range of TTP) is the same regardless of whether the risk elevates to 110 or 120 in terms of FLC level. In other words, even if the ML algorithm predicts the TPP within 10% accuracy, whereas the FLC predicts it to within 20%, there is no change in the treatment……."

"…..there seems to be enough evidence to assign FLC levels to risk……." The risk determines the treatment (not whether the risk is calibrated to a specific value in a range).

"…….100 days, you generated multiple confusion matrices by varying the classification threshold and calculated AUC on that day for circulation and non-circulation events?…." i did not understand the answer. Did you vary the classification thresholds for each AUC point (100 days) If so, how was this actually done?

"…….Hence, what you are essentially saying in Figure 5 is that the (relative) inaccuracy of the model to ID DDTOC can be used to differentiate between circulatory deaths and DDTOC. I do not think this is a valid scientific proposition."… you do suggest this in the manuscript but do not provide a valid answer.

Briana Munteanu <brianamunteanu1@gmail.com>
Sun, Nov 2, 12:45 PM
to me

Dear Dr. Shireesh Apte,

Thank you very much for your valuable comments on our manuscript. We appreciate the time and effort you have taken to provide constructive feedback. Please find below our detailed responses to each comment. We hope that these revisions adequately address the concerns raised.

1.The reviewer is concerned that your model with a 25% false negative rate and a similar false positive rate does not improve on the binary cut off for classification of risk as 'low' or 'high'. They are not asking for lumping your individual ML results. Can you match each of your individual results with whether those individual results correctly fall into the literature binary model ?

We would like to thank the Editor for clarifications. We could not identify any published recommendations defining specific cutoff values of FLC that would guide pharmacological management for circulatory diseases or for the aggregate of all-other conditions, as presented in the original study by Dispenzieri et al. (2012). To facilitate comparison between our machine learning (ML) model and the standard binary classification approach, we examined patients in the top decile of FLC levels - those identified in the original research as being at the highest risk of mortality.

Among these patients, 61% died during the observation period, meaning that the conventional cutoff-based approach would classify approximately 39% of survivors as “high risk.” In contrast, our ML model correctly identified 75% of the individuals from this high-FLC group who actually died during the study. These findings suggest that the ML approach provides a more accurate, individualized risk assessment than the binary cutoff method, supporting its potential utility as a complementary tool in medical decisions.

In light of this, we would appreciate your guidance on whether the following explanation added to the Discussion section would be acceptable for the reviewer:

“We compared the machine learning (ML) model with the conventional binary FLC cutoff by focusing on patients in the top decile of FLC levels, identified as highest-risk in previous work (Dispenzieri et al., 2012). Among these patients, 61% died during follow-up, meaning the standard approach would misclassify roughly 39% of survivors as “high risk.” In contrast, the ML model correctly predicted 75% of deaths within this group, highlighting its ability to provide more precise, individualized risk assessments and complement traditional methods in clinical decision-making.”

2 and 3. The reviewer is asking that since your model's results do not affect subsequent pharmacology; then what is the use of a 'greater' discriminatory value provided by your model? If (say) your model predicted greater risk so that a different pharmacotherapy could be applied, then that discriminatory nuance would have some value. If it does not, why is a 'complementary approach to classical statistical analysis' needed? If you can find literature that suggests that dose-titration or different therapy regimens or protocols in high risk patients improve survival; that would be a way to answer this question. If you cannot find such literature, you can put this down as a limitation of your approach.

By “complementary approach,” we refer to the ML model’s ability to integrate all available patient data to generate personalized survival predictions, rather than simply grouping patients in two (or more) groups by FLC levels. Since these predictive methods have only recently begun to be explored (Durso-Finley J. et al., 2024), we found no medical studies addressing personalized care or pharmacological regimens tailored to the fine granularity of FLC levels proposed here.

Durso-Finely J., Barile B., Falet J-P, Arnold D.L., Pawloski N., Abel T. (2024), “Probabilistic Temporal Prediction of Continuous Disease Trajectories and Treatment Effects Using Neural SDEs”, https://arxiv.org/pdf/2406.12807

Based on your suggestions, we propose to add the following paragraph in the Limitation section:

“A limitation of this study is that, although the ML models integrate comprehensive patient data to generate individualized survival predictions, such predictive approaches are still in their early stages of development (Durso-Finley J. et al., 2024). Consequently, no existing medical studies support the application of personalized care or pharmacological regimens based on the fine granularity of FLC levels proposed in this work.”

4. Your paragraph is acceptable. However, I still would like to see a different nomenclature than "AUC". Can you confirm that Hung & Chiang, 2010, used this same nomenclature "dynamic AUC" in their work? If yes, then it is acceptable. If not, then please use the term that they used.

We thank the Editor for this helpful comment and for the opportunity to clarify the terminology. Huang and Chiang originally introduced the concept of dynamic AUC, and their methodology has since been implemented in widely used statistical packages for both Python and R. A Google Scholar search for “dynamic AUC” identified approximately 489 publications applying this

approach, demonstrating its broad adoption in the research community. The corresponding search results can be accessed at the following link:

https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=%22dynamic+AUC%22&btnG=

5. Does the AUC calculation for circulatory related deaths use different (or a subset) of features than the AUC calculation for all-cause mortality? If yes, then you have a different model for these two conditions and you will need to state that. If not, the paragraph that you have proposed is acceptable.

Thank you for your insightful comment. We confirm that both the circulatory-related mortality model and the all-cause mortality model use the same set of input features. The difference between models is made by using different targets associated with the different outcomes.

We sincerely thank you again for your thoughtful comments and guidance.

Sincerely,

Briana Munteanu

Otilia Barbu, PhD

shireesh apte
Sun, Nov 2, 8:37 PM
Dear authors, All your responses are acceptable. However, for point 1, please put down a caveat of this being only one study that was compared. Please revise an

Briana Munteanu <brianamunteanu1@gmail.com>
Wed, Nov 5, 9:41 AM
to me

Dear Dr. Shireesh Apte,

Thank you for your feedback and guidance. We have revised the manuscript and resubmitted it. Specifically, we have added the following caveat in response to point 1 in the Conclusions section: "However, these findings are based on a single comparative study and should be interpreted cautiously until validated in larger, independent cohorts."
We appreciate your time and thoughtful review.

Thank you,
Briana Munteanu
Otilia Barbu, PhD

Briana Munteanu <brianamunteanu1@gmail.com>
Fri, Nov 7, 5:47 AM
to me

Dear Dr. Shireesh Apte,

Thank you for reviewing our manuscript and the previous reviewer comments. We write to clarify the matter and to seek the Editor's guidance on whether our decision to address the reviewers' concerns without adding a paragraph to the manuscript is appropriate. Below, we provide our justifications for this decision.

there is data leakage by using 'chapter' (cause of death) as a predictor when outcomes are defined from it; this makes the models non-deployable at baseline and inflates performance.
We respectfully believe that the variable chapter, which is only used to construct the necessary targets for our models, could not leak into the models.

First, the variable chapter is not used among the features used for training. We refer below to the paragraph from the Data Preprocessing section, describing how the target variables were derived from the original target chapter variable:

"The target variable was defined according to each objective, starting from the original chapter data. For the first objective, two distinct categories were defined: "censored", representing patients with unknown survival outcomes due to right censoring, and "died", including all patients who passed away during the study period. For the second objective, the target variable defined the following two categories: the "circulatory" class, of patients who died from circulatory-related conditions, and the "other" class containing all remaining patients, including both right-censored individuals and those who died from other causes."

Second, when dividing the original data in 'training' and 'testing' groups, no leakage did happen between these randomly selected groups. All the results that we have reported come from the 'testing' group, in which the 'chapter' values were only used to compare the models' predictions with the known results.

We respectfully request that the Editor take this clarification into account and recognize that chapter was not treated as a predictor in our analysis. We hope this resolves any potential misunderstanding regarding the role of chapter in our study.

Although tree methods mitigate collinearity, no empirical checks (correlation matrices, VIF) were reported; SVM remains vulnerable to redundant features without proper preprocessing.
In most published manuscripts employing ML models such as SVM, Random Forest (RF), or XGBoost, authors do not typically report correlation matrices or Variance Inflation Factors (VIF) because: (1) VIF is specifically designed for linear regression models, where multicollinearity directly inflates coefficient variance and affects interpretability; it is not generally applied to SVM, RF, or XGBoost, as these algorithms can internally handle redundant features (as discussed in our previous response); and (2) correlation matrices are primarily exploratory diagnostics used to examine feature relationships, but they are not a methodological requirement for nonlinear, tree-based, or kernel-based models.

However, to address the reviewer's request, we have computed the VIF

image.png

and the correlation matrix

image.png

We did not include these results in the revised manuscript, as reporting them is not standard practice for ML algorithms. We kindly request the Editor's approval for this decision.

LabelEncoder on categorical variables for SVM imposes spurious ordinality; one-hot encoding is required.
Two categorical variables - sex and chapter - were encoded using the LabelEncoder in our analysis.

The variable chapter served as the target, and using label encoding for the target does not impose spurious ordinality in SVMs; this is standard practice in classification tasks (Bisong, 2019). The issue of artificial ordinality arises only when label encoding is applied to categorical predictor variables. One-hot encoding increases the dimensionality of the target unnecessarily, which is not needed for these algorithms and can complicate downstream computations like metric evaluation.

The variable sex is strictly binary (male, female), so LabelEncoder and OneHotEncoder convey equivalent information, yielding numerically identical results after model fitting. However, if sex includes more than two categories (e.g., male, female, unknown), one-hot encoding would be necessary to avoid spurious ordinality and maintain model generalizability.

Accordingly, we have opted not to include a paragraph on this topic in the manuscript, but we would be grateful for the Editor's advice.

Bisong, E. (2019). Introduction to Scikit-learn. In Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners (pp. 215-229). Berkeley, CA: Apress.

Thank you,

Briana Munteanu

Otilia Barbu, Phd

shireesh apte <journalofhighschoolscience@gmail.com>
Sun, Nov 9, 11:18 PM
to Briana

Dear author,

1. I do not agree with you on point 1. In objective 2, the label 'circulatory' vs 'other' is derived directly from cause of death information (chapter), which is not available at baseline. This means that to assign labels for training, you already need to know the cause of death (i.e. the future outcome). Hence, chapter is still post-outcome information, even though it may not be input per se'. Thus, the labels themselves are defined using data that would not be available at prediction time. This is still target leakage - even though it may not be construed as feature leakage, and it still leads to optimistic, non-deployable models, because you are teaching the model to predict something that is already derived from future information.

2. Please include a sentence to the effect that 'multicollinearity was assessed and all feature VIF were less than 3.25" or something similar. That will suffice.

3. Please justify briefly in the manuscript using "The variable chapter served as the target, and using label encoding for the target does not impose spurious ordinality in SVMs; this is standard practice in classification tasks (Bisong, 2019)." and include the reference you mention in the manuscript.

Please address all concerns in the platform going forward. I will include these in the review and send out another review and revise request, although point 1 seems non-salvageable.
Best,
Shireesh Apte

Dear author,
I have asked another reviewer, who is an expert in the AI/ML field, to review your paper. Please be patient until they provide a review.
Best,
Shireesh Apte

•Shireesh Apte
Dec 9, 2025 - 9:33 pm IST

Dear author,
We are waiting on the reviewer. If we don't hear back this week, we will make a decision. I apologiize for the delay and appreciate your patience.
Best,
Shireesh Apte

•Shireesh Apte
Dec 10, 2025 - 10:00 am IST

Dear author,
The reviewer has responded that your logic is correct. Hence, we will proceed with your manuscript. I will issue a decision of accept now and we will begin copyediting.
Best,
Shireesh Apte