A Deep Learning Ordinal Classifier

Tiphelele Lwazi Nxumalo[1], Richard Maina Rimiru[2], Vusi Mpendulo Magagula[3]
Department of Mathematics, Pan African University Institute for Basic Sciences, Technology and Innovation (PAUSTI), Nairobi, Kenya[1]
School of Computing and Information Technology (SCIT), Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya[2]
Department of Mathematics, University of Eswatini, Matsapha, Eswatini[3]

***Abstract*—Deep learning models such as TabNet have gained popularity for handling tabular data. However, most existing architectures treat categorical variables as nominal, ignoring the inherent ordering in ordinal data, which can lead to suboptimal classification performance, particularly in tasks where ordinal relationships carry meaningful information, such as quality assessment, disease severity staging, and risk prediction. This study investigates the impact of explicitly modeling ordinal relationships in deep learning by developing an ordinal classification model and comparing it with its nominal counterpart. The proposed approach integrates TabNet a deep learning framework with ordinal constraints, leveraging a proportional odds model to better capture the ordinal structure and Beta cross-entropy as the loss function to enforce ordering during training. To evaluate the effectiveness of the proposed ordinal classification approach, experiments were conducted on two publicly available datasets: the White Wine Quality dataset and the Hepatitis C dataset. The results demonstrate that incorporating ordinal constraints leads to improvements across multiple evaluation metrics, including 1-off accuracy, average mean absolute error (AMAE), maximum mean absolute error (MMAE), and quadratic weighted kappa (QWK) compared to a nominal classification model trained under the same conditions. These findings underscore the importance of ordinal modeling in tabular classification and contribute to the advancement of deep learning techniques for structured data.**

Keywords*—Ordinal classification; TabNet; proportional odds model; tabular data

I. Introduction

Tree-based machine learning algorithms like Extreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost) have achieved strong performance on tabular data, but they have limitations in learning complex, and non-linear relationships as compared to deep learning methods. Numerous neural architectures have been proposed for the purpose of strengthening neural networks' performance on tabular data. TabNet [1] is a type of neural network specifically designed for processing tabular data. TabNet has improved classification in various domains such as insurance [2], rainfall prediction [3], food safety risk [4]. In tabular datasets, each column represents a distinct feature, with some columns containing continuous numerical values while others include discrete or categorical data [5]. During training, TabNet uses softmax for discrete outputs which gives the model's predefined set of class probabilities for classification tasks. However, on the case of ordinal classification, the softmax might not be the best choice.

Ordinal regression (ordinal classification) problems in machine learning involve classifying patterns according to a categorical scale that reflects a natural order among the labels [6]. This type of problem can be approached as nominal classification; however, doing so ignores the ordinal information [7], which may result in low prediction accuracy and the loss of important information regarding the order of the categories. A more effective strategy is to employ methods that consider the ordinality, thereby enhancing the classification model's performance. It can be challenging to ascertain the link between distinct classes using other techniques, but ordinal regression can help [8].

In non-tabular domains, ordinal classification has been transformed by deep learning, such as age estimation [9] and medical diagnosis [10] using images. However, no deep learning model has been developed explicitly for ordinal classification in tabular data. This study intends to close this gap by creating a deep learning ordinal classifier specifically designed for tabular data, utilising neural networks with ordinal constraints to enhance interpretability and prediction accuracy. We introduce Proportional Odds Model (POM) for TabNet, combined with the Beta Cross-Entropy loss function, to enhance the classification performance of ordinal tabular data. The (POM) [11] is a category of generalized linear models employed to model the dependence of an ordinal response on discrete or continuous covariates. The POM can be directly applicable to the output of a TabNet, thus addressing the challenge of deep learning methods in tabular data ignoring ordering information of data. POMs offer a more adaptable and comprehensible method of deep ordinal classification by indirectly modelling a latent space in addition to the set of thresholds dividing the ordered classes. By replacing the one-hot labels with their soft label equivalents, the beta cross-entropy loss function adds soft labels to the cross-entropy loss function. Soft labels might potentially improve model performance by better accounting for ordinal classification uncertainty, which occurs when it is difficult to distinguish between nearby categories because of their resemblance.

The remainder of this paper is structured as follows: A review of relevant theory and related literature is presented in Section II; materials and methods for completing the work are described in Section III; analysis and interpretation of results are presented in Section IV, while Sections V and VI provide the discussion and conclusion, respectively.

II. Literature Review

While a lot of research has been done on the ordinal classification of tabular data, very little of it has concentrated on deep learning for the ordinal classification of tabular data. Convolutional Neural Networks (CNN) are used in image datasets for the current deep learning ordinal techniques.

A. Deep Learning Ordinal Classification in Image Data

For determining the degree of neurological damage in individuals with Parkinson's disease (PD), an ordinal decomposition method in conjunction with a 3D CNN ordinal model was suggested [10]. Instead of employing a softmax function for the output nodes, a regular sigmoid function is supplied in the output node. They provided experimental evidence that using ordinal information can enhance performance on a challenging task, such as evaluating changes in brain activity in Parkinson's disease.

By taking into account a family of probabilistic ordinal link functions in the output layer, a deep convolutional neural network model for ordinal regression was proposed [9]. The experiments ran over two different image data ordinal classification problems. The link functions used are those from cumulative link models, which are traditional statistical linear models that project each pattern onto a one-dimensional space.

B. Ordinal Classification in Tabular Data

A thorough analysis of ordinal classification techniques was presented in study [6], the authors grouped ordinal classification methods into three: naïve approaches, binary decomposition, and threshold models. Naïve approaches apply standard machine learning models without explicitly considering the ordinal structure. Binary decomposition transforms the ordinal problem into multiple binary classification tasks, either solved by separate models or a multi-output model. Threshold models approximate a real-valued predictor and partition it into intervals to determine class boundaries.

In naïve approaches, artificial intelligence-machine learning (AI-ML) algorithms were proposed for cost-sensitive learning utilizing resampling techniques and for ordinal categorization using ordinal decomposition [12]. They evaluated a "naïve" multi-class decomposition called "One-Vs-One" (OvO) and a "naïve" conversion of the classification issue into a regression task, and an ordinal 'Ordered Partitions' (OrdP) decomposition. In the cost-sensitive learning they used SMOTE. To predict white wine quality based on physicochemical data, [13] applied Synthetic Minority Oversampling Technique (SMOTE) algorithm to address class imbalance then applied Random Forest and Multinomial Logistic Regression for classification, ignoring the order between classes. Random Forest outperformed the Multinomial Logistic Regression. The absence of a clear correlation between the regression model's prediction error and the misclassification error is one of the drawbacks of the conventional ordinal classification techniques based on regression.

An ordinal binary decomposition method that allows ordering information to be used by standard classification in class attributes was presented in study [14]. An ensemble-based classifier that combines ensemble-learning paradigm such as bagging and AdaBoost with the ordinal binary decomposition by study [14] to improve prediction performance was proposed in study [15]. To predict soil temperature level, the study in [16] proposed Soil Temperature Ordinal Classification (STOC) approach that used five different traditional ML methods (K-Nearest Neighbors, Random Forest, Naïve Bayes, Support Vector Machines, and Decision Trees). The STOC using Decision Trees as the base learner (STOC.DT) performed better among the others. The primary challenge with ordinal binary decomposition approaches is that, they are strongly dependent on the specific decomposition method used and the way the results from all decompositions are combined into a final classification.

Two gradient descent-based techniques for learning an ensemble of base classifiers being decision rules was presented in study [17] . The forward stage-wise additive modelling that makes use of the threshold loss function is the foundation of the decision rule induction algorithm. The ordinal decision criteria are competitive with both the established ordinal classification techniques and conventional regression and multi-class classification methods. In study [18] a method that simplifies the ordered class classification problem to the conventional two-class problem was presented. Neural networks and support vector machines were then trained using the method. An experimental study verified the usefulness of the approach. In study an ordinal loss function based on the soft labelling approach was used to combine four Multi-Layer Perceptron (MLP) models that had been optimized. Furthermore, an ordinal logistic regressor is included with the soft labelling models. The unimodal probability distributions fail to explicitly model the ordinal structure of data.

C. Unimodal Regularisation

The performance of ordinal classifiers with respect to the conventional one-hot encoding has been enhanced by the distributions suggested to softly model the targets.

A straightforward technique was proposed in study [20] to enforce unimodality in discrete ordinal probability distributions using the Poisson distribution. The distribution parameter λ is equal to the mean and variance of this type of distribution. As a result, its ability to obtain a slight variation is limited. Because of this, they also employed the binomial distribution, which has two parameters: the probability, p, and the number of classes, C. Although the variance ($Cp(1-p)$) and the mean (Cp) have different expressions, positioning the mode at the right point in the interval while obtaining a small variance is difficult.

It was suggested to use a soft labelling strategy based on generalized triangular distributions, which are asymmetric and unique for every class in study [21]. A metaheuristic is used to calculate the parameters of these distributions, which are then tailored to the particular problem. Additionally, the model can avoid errors in remote classes thanks to this method.

A sample based on the exponential function $e^{\frac{-|i-l|}{\tau}}$ where l represents the class of the pattern and $i = 1, \ldots, C$, followed by a softmax normalization was proposed [22]. However, the value of τ requires experimental tuning, and in some cases, the probability mass is not sufficiently concentrated in the interval of the correct class.

A unimodal regularization technique based on the beta distribution was proposed in study [23] and applied to the cross-entropy loss. This regularization encourages the label distribution to form a soft unimodal shape. Because of its low variance and domain constraint from 0 to 1, using beta distributions to determine the soft labels is an improvement over earlier approaches [19] .

D. Research Gap and Motivation

Ordinal binary decomposition (OBD) is commonly used to handle ordinal classification in tabular data. OBD does, however, have inherent limits because its effectiveness is highly reliant on the particular decomposition technique employed and how the output of several decompositions is combined to provide a final classification. This dependence may result in suboptimal performance and a more complex model. To address these challenges, we propose an alternative approach inspired by techniques widely used in image-based ordinal classification namely, threshold-based modeling applied to the output of deep learning algorithms. We use TabNet [1], a deep learning model developed especially for tabular datasets, and apply POM to its output layer.

Additionally, recent research has shown that soft labeling can improve ordinal classification performance by incorporating uncertainty and reducing the impact of hard class boundaries. To take advantage of this benefit, we use a unimodal regularization technique based on the beta distribution [23] in place of the conventional categorical cross-entropy loss in order to improve the accuracy and robustness of our ordinal classifier.

In order to provide a more efficient solution for ordinal classification in tabular data, our study aims to close the gap between conventional OBD approaches and contemporary deep learning techniques by using these developments.

III. Materials and Methods

Building on the previous analysis of the state-of-the-art, our proposal is to integrate a flexible threshold model in the output layer, POM, with a unimodal probability distribution based on the beta distribution to more effectively enforce ordinal constraints during learning.

A. Data Description and Preprocessing

This study uses two datasets to evaluate the different models; Hepatitis C dataset and white wine quality dataset both obtainable online at UCI machine learning repository [24]. The data was processed and split into the ratio of 7:3 for training, and testing respectively.

1) Hepatitis C dataset: The Hepatitis C dataset has 615 instances of laboratory values of blood donors and Hepatitis C patients and demographic values like age. It includes a total of 14 features including the target attribute which has five outcomes, ‘0=Blood Donor’, ‘0s=suspect Blood Donor’, ‘1=Hepatitis’, ‘2=Fibrosis’, ‘3=Cirrhosis’. Category (blood donors vs. Hepatitis C, including its progression: 'simply' Hepatitis C, Fibrosis, Cirrhosis) is the target attribute for classification. The dataset has some missing values and they were filled using mean. Blood donor, suspect blood donor was encoded as 0, hepatitis was encoded as 1, fibrosis encoded as 2, cirrhosis as 3. Numerical values were normalized. Since the classes were imbalanced (see Fig. 1), SMOTE was used to balance the classes.

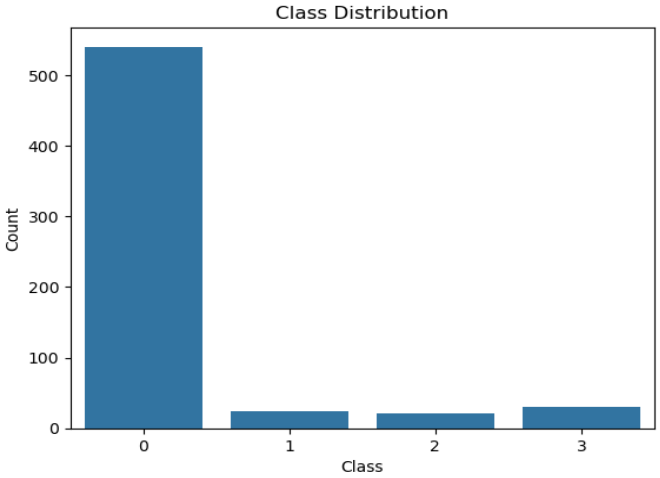

Fig. 1. Hepatitis C dataset class distribution.

2) White wine quality dataset: The white wine quality dataset has 4898 instances of physicochemical tests of the Portuguese “Vinho Verde” wine. It includes a total of 12 features including the target variable “quality” which has 7 outcomes ranging from 3 to 9. The classes are ordered and not balanced as shown in Fig. 2 so SMOTE was used to balance the classes.

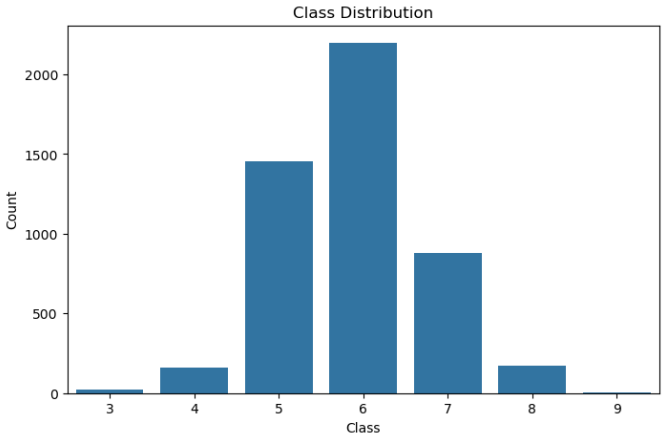

Fig. 2. White wine quality dataset class distribution.

B. TabNet Architecture

TabNet's architecture consists of N_{steps} subnetworks that are processed sequentially in a hierarchical manner (see Fig. 3), with each subnetwork representing a decision step. During training, every decision step processes the current data batch as its input. At the i^{th} step the subnetwork takes in the processed information from the $(i-1)^{th}$ step to determine which features to utilize. It then outputs a refined feature representation, which is incorporated into the overall decision. TabNet combines the outputs of all decision steps to generate the final prediction.

At every decision step, TabNet employs a feature mask that encourages controlled sparsity $M[i] \in \Re^{B \times D}$, where B

represents the batch size, for soft instance-wise feature selection. The masking is applied multiplicatively, $M[i] \cdot f$, f is the feature representation at the current step. This feature mask is learned using attentive information from the preceding decision step, $a[i-1]$, and is computed as: $M[i] = \text{sparsemax}\left(P[i-1] \cdot h_i(a[i-1])\right)$. The feature transformer module determines which features should be forwarded to the next decision step and which features should be utilized to produce the output at the current decision step. This process is defined as: $[d[i], a[i]] = f_i(M[i] \cdot f)$, where $d[i] \in \Re^{B \times N_d}$ represents the decision step output, and $a[i] \in \Re^{B \times N_a}$ serves as attentive information for subsequent steps. Certain layers within the feature transformers are shared across all decision steps. The feature masks generated during this process correspond to local feature weights and can be aggregated into a global importance score.

Drawing inspiration from decision-tree-like aggregation, TabNet forms the overall decision embedding as: $d_{out} = \sum_{i=1}^{N_{steps}} \text{ReLU}(d[i])$. A linear transformation, $W_{final} d_{out}$, is then applied to generate the output mapping. For discrete outputs, a softmax function is used during training, while argmax is applied during inference.

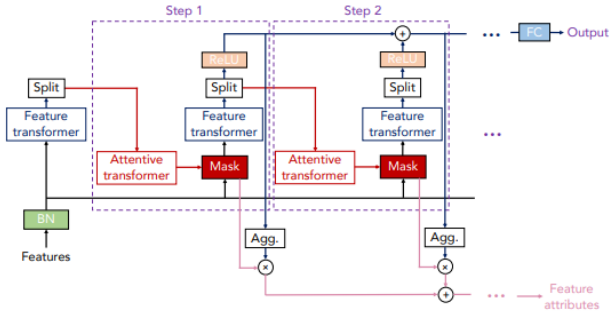

Fig. 3. TabNet architecture [1].

C. Proportional Odds Model

When the classes have a natural order, rather than addressing the problem using the standard approach mentioned above, a threshold-based method known as the Proportional Odds Model (POM) can be used instead of softmax. POM is part of a broader category of models called Cumulative Link Models (CLMs) [25]. In the POM framework, the class ordering is maintained through the following latent constraint shown in Eq. (1):

$$f^{-1}(P(y \le y_c|x)) = t_c - f(x) \quad (1)$$

Where $c = 1, 2, \cdots, C-1, f^{-1}$ is a function that maps probabilities from the range [0,1] to the entire real number line, ensuring a monotonic transformation. The threshold for class y_c is denoted as t_c. Consequently, the class y_c is predicted if and only if: $f(x) \in [t_{c-1}, t_c]$.

POM utilizes the logit link function, which is defined in Eq. (2) as:

$$\text{logit}[P(y \le y_c|x)] = \log \frac{P(y \le y_c|x)}{1 - P(y \le y_c|x)} = t_c - f(x), \quad c = 1, \cdots, C-1, \quad (2)$$

or the equivalent expression expressed in Eq. (3):

$$P(y \le y_c|x) = \frac{1}{1 + e^{-(t_c - f(x))}} \quad (3)$$

D. Beta Cross-Entropy

Beta cross-entropy is a unimodal regularization technique that incorporates the beta distribution into the cross-entropy loss. This regularization promotes a soft unimodal distribution of labels, making it more suitable for ordinal classification problems.

For a one-hot label, the probability distribution of the label is given by $q(i) = \delta_{i,l}$, where l represents the ground truth class. The Dirac delta function, $\delta_{i,l}$ equals 1 when $i = l$, and 0 otherwise. This label smoothing technique can be incorporated into the cross-entropy loss by modifying $q(i)$ in Eq. (4):

$$L = \sum_{i=1}^{J} q(i)[-\log P(y = C_i|x)] \quad (4)$$

with a target distribution that is more conservative as shown in Eq. (5):

$$L = \sum_{i=1}^{J} q'(i)[-\log P(y = C_i|x)] \quad (5)$$

where $q'^{(i)} = (1-\eta)\delta_{i,1} + \eta f(x,a,b)$ and the linear combination is controlled by the parameter η. $f(x,a,b)$ represents the probability value sampled from a beta distribution centred in $x = \frac{2J-1}{2J}$ and makes uses of the a and b parameters obtained using the method proposed by the authors [23].

The properties of the beta distribution are as follows. In its standard form, the beta distribution, denoted as, $\beta(a,b)$ is a continuous distribution. Its probability density function (PDF) is given in Eq. (6):

$$f(x,a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \quad (6)$$

where $0 < x < 1, a > 0$ and $b > 0$. The beta function $B(a,b)$ has the form shown in Eq. (7):

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad (7)$$

where $\Gamma(a) = (a-1)!$. When $a, b > 1$, the probability density function $f(x)$ has a unique mode at $\frac{a-1}{(a+b-2)}$ and is zero at $x = 0$ and $x = 1$. If $a = 1$ or $b = 1$ then $f(x)$ has a corresponding terminal value b or a, respectively. Lastly, $f(x)$ becomes the uniform distribution if $a = b = 1$.

Fig. 4 illustrates the differences in the final layer and loss functions of the nominal TabNet and its ordinal variation as proposed.

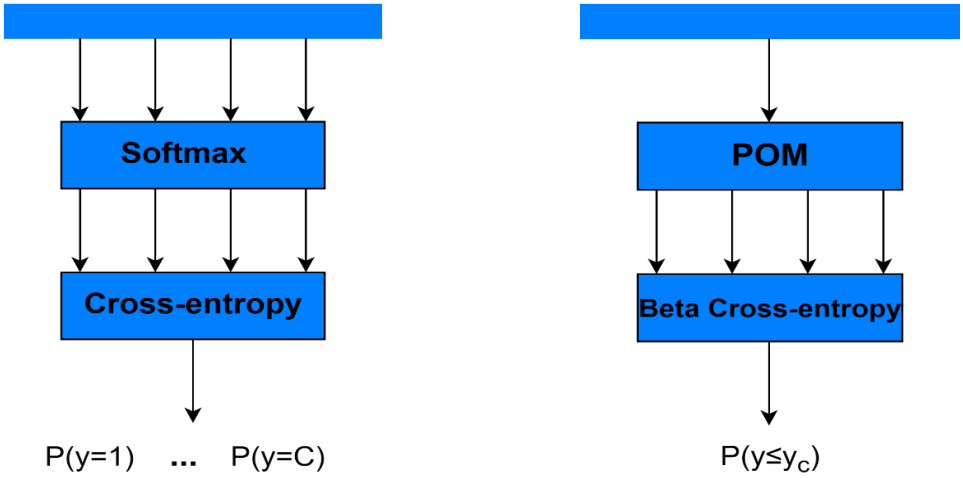

Fig. 4. Comparison between the existing nominal (left) and proposed ordinal TabNet (right). The key difference is in the loss function and the constraint on learned representations, affecting how the model treats ordinal relationships.

IV. Results

The results of the proposed ordinal TabNet approach are presented in this section, along with a comprehensive comparison against the compared approaches.

A. Hyperparameters

We present best hyper-parameter configuration that achieved the highest performance for each dataset. We employed Bayesian hyper-parameter optimization approach to identify the most effective hyper-parameter setup for optimization purposes. We used early stopping as a strategy to determine the optimal number of epochs for training the model, which helps conserve computational resources, prevent over-fitting, and demonstrate strong generalization capabilities without excessive training.

For the Hepatitis C dataset, the best performance was achieved with the values shown in Table I.TABLE I.

TABLE I. Hyperparameters for Hepatitis C Dataset

Hyperparameter	Value
Number of Decision Steps (n_steps)	7
Decision Layer Size (n_d)	39
Attention Layer Size (n_a)	31
lambda_sparse	2.882 x 10^{-3}
Learnring rate (lr)	7.457 x 10^{-3}
Gamma	1. 175

For the white wine dataset, the best performance was achieved with the values shown in Table II.

TABLE II. Hyperparameters for White Wine Datatset

Hyperparameter	Value
Number of Decision Steps (n_steps)	8
Decision Layer Size (n_d)	62
Attention Layer Size (n_a)	63
lambda_sparse	3.634 x 10^{-3}
Learnring rate (lr)	9.890 x 10^{-3}
Gamma	1. 010

B. Evaluation Metrics

Various evaluation metrics are used to measure the closeness of predictions to actual values. In this work, all selected performance metrics are well-suited for ordinal classification problems, as they appropriately penalize misclassification errors more severely when they occur in distant classes compared to adjacent ones. The following performance metrics are considered:

- 1-off accuracy: assesses the proportion of predictions that are either correct or differ by at most one category from the actual class.
- Average Mean Absolute Error (AMAE) [26]: The average MAE, calculated as the mean of the MAE classification errors across different classes, helps to reduce the impact of imbalanced class distributions. When AMAE is applied to an unbalanced dataset, the trivial class for AMAE is counted like any other class rather than in proportion to its frequency. Let MAE_c be the MAE for a given c-th class, AMAE is defined in Eq. (8) as:

$$AMAE = \frac{1}{C}\sum_{c=1}^{C} MAE_c \quad (8)$$

where AMAE values fall between 0 to $C-1$.

- Quadratic Weighted Kappa (QWK) [27]: Reflects the degree of disagreement, placing greater emphasis on larger differences between ratings than on smaller ones. The quadratic weighted kappa is calculated as Eq. (9):

$$QWK = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \quad (9)$$

where, W is the penalization matrix; quadratic weights are taken into consideration in this instance, $W_{i,j} = \frac{(i-j)^2}{(C-1)^2}$), E is the expected matrix, whereas O is the confusion matrix that represents the agreement that would occur by chance.

- Maximum Mean Absolute Error (MMAE) [28]: MMAE represents the MAE value of the class with the largest

deviation between the true and predicted values, as shown in Eq. (10):

$$MMAE = \max\{MAE_c; c = 1, \cdots, C\} \quad (10)$$

C. Compared Approaches

The proposed ordinal TabNet approach is evaluated in comparison with the following methods:

- A nominal TabNet (using softmax and cross-entropy) [1].
- STOC.DT [16]: An ordinal classification model that was developed to classify soil temperature level in tabular data.

D. Model Comparison

This section presents the results of this study that implemented the ordinal TabNet. Table III and Table IV present a comparative analysis of the proposed approach against the baseline nominal model TabNet, and STOC.DT using evaluation metrics for both the Hepatitis C and white Wine datasets. Each metric's best value is indicated in bold.

TABLE III. HEPATITIS C MODEL EVALUATION METRICS

Model	1-off (%) ↑	AMAE ↓	QWK ↑	MMAE ↓
TabNet	97.8	**0.423**	0.835	0.777
STOC.DT	97.2	0.602	0.769	1.16
Proposed Approach	**98.9**	0.439	**0.890**	**0.666**

TABLE IV. WHITE WINE MODEL EVALUATION METRICS

Model	1-off (%) ↑	AMAE ↓	QWK ↑	MMAE ↓
TabNet	92.6	1.222	0.584	3.0
STOC.DT	92.2	**1.028**	0.569	**2.16**
Proposed Approach	**92.9**	1.051	**0.598**	2.333

Test confusion matrices for the Hepatitis C and white wine datasets are displayed in 0 and Fig. 6 , respectively, for the proposed approach and the baseline approach (nominal approach).

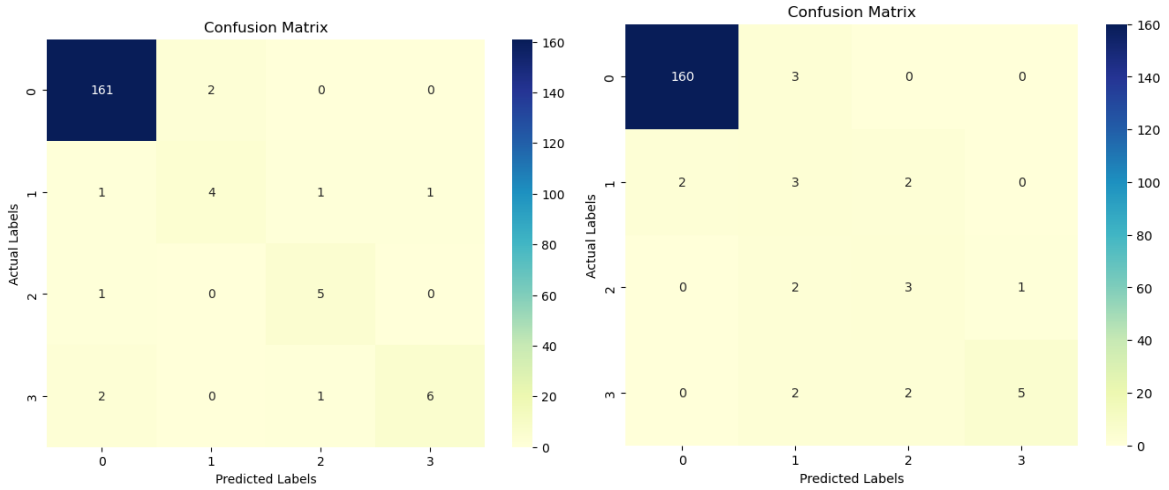

Fig. 5. Hepatitis C confusion matrices for nominal(left) and proposed ordinal TabNet(right).

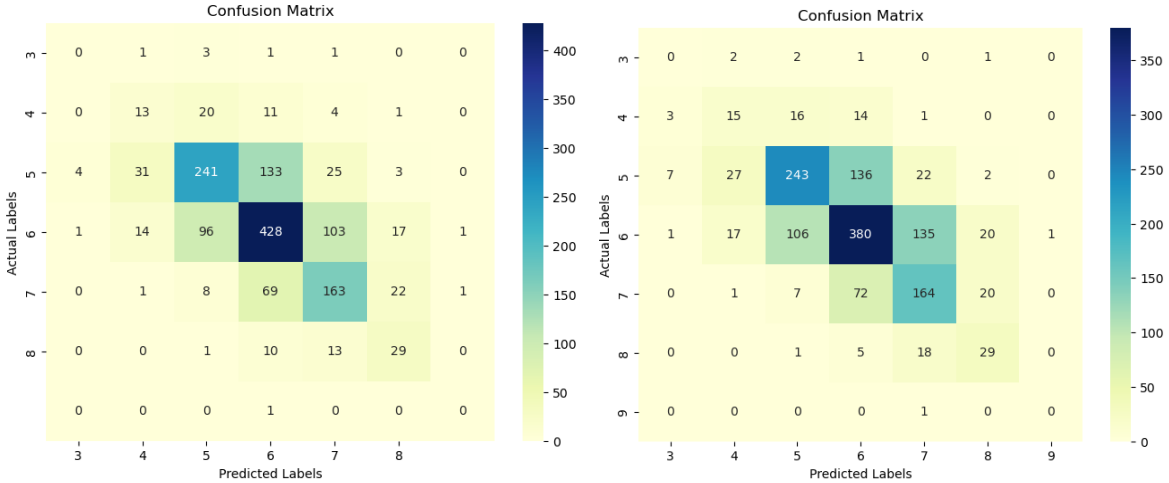

Fig. 6. White wine confusion matrices for nominal(left) and proposed ordinal TabNet(right).

V. Discussion

Table III shows results from Hepatitis C model evaluation, our proposed approach has achieved a 1-off accuracy of 98.9%, QWK of 0.890, and MMAE of 0.666 outperforming TabNet [1] that treats the problem as nominal and STOC.DT [16] that takes the ordinal information into consideration through ordinal binary decomposition. When comparing the test confusion matrices (0) of the baseline technique (nominal approach) and proposed approach, the confusion matrix of the proposed approach is centered on the diagonal which shows that our approach penalizes inaccuracy among distant classes.

Table IV shows results from white wine model evaluation. It demonstrates that, in comparison to the alternative methods, TabNet [1] and STOC.DT [16], the proposed method achieved a higher 1-off accuracy and QWK with values 92.9% and 0.598, respectively. STOC.DT was a close competition as it performed slightly better than our approach in terms AMAE and MMAE. The same can be observed for white wine test confusion matrices Fig. 6 as in the Hepatitis C confusion matrices that the confusion matrix of the proposed approach is centered on the diagonal which shows that our approach penalizes inaccuracy among distant classes.

VI. Conclusion

This paper presents a novel deep ordinal network that integrates POM with a Beta Cross-Entropy loss function applicable to ordinal tabular data. The study presents a data-driven approach to improving predicting accuracy while preserving the inherent order within categorical labels by combining deep learning architecture with ordinal constraints. The proposed model enhances the performance of deep networks compared to its nominal counterpart. The findings indicate that the optimal parameter values are problem-dependent, emphasizing the need for an experimental design where all parameters are carefully tuned for each specific problem.

By emphasizing the benefits of integrating ordinal constraints into deep neural networks, this paper theoretically advances the expanding field of ordinal deep learning. Additionally, the study provides insight into how deep ordinal classifiers behave while working with tabular data, laying the groundwork for further developments in this field.

The proposed approach can successfully classify ordinal data with enhanced robustness, which makes it appropriate for practical applications where ordinal relationships are essential.

Despite these contributions, the study has certain limitations. Substantial computational resources are needed for the deep learning model, which restricts its use in real-time situations. The model's effectiveness on other ordinal classification tasks has not been tested, despite its strong performance on the selected datasets.

Future work could explore an ensemble approach that integrates various soft labeling techniques to enhance model robustness. Additionally, investigating alternative cumulative link model (CLM) link functions beyond the logit function may provide deeper insights into ordinal relationships and improve classification performance.

Acknowledgment

The authors express their gratitude to Pan African University, Institute for Basic Sciences, Technology and Innovation (PAUSTI) for their financial contribution and to the Department of Mathematics for their unwavering support in the completion of this work.

References

[1] S. Ö. Arık and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in Proceedings of the AAAI conference on artificial intelligence, 2021, pp. 6679-6687.

[2] M. Kevin, M. Finbarr, S. Barry, M. Leandro and C. German, "Deep learning in insurance: Accuracy and model interpretability using TabNet," Expert Systems with Applications, vol. 217, p. 119543, 2023.

[3] Y. Jianzhuo, X. Tianyu, Y. Yongchuan and X. Hongxia, "Rainfall forecast model based on the tabnet model," Water, vol. 13, p. 1272, 2021.

[4] Y. Chen, H. Li, H. Dou, H. Wen and Y. Dong, "Prediction and visual analysis of food safety risk based on tabnet-gra," Foods, vol. 12, p. 3113, 2023.

[5] J. A. Marais, "Deep learning for tabular data: an exploratory study," Stellenbosch University, Stellenbosch, 2019.

[6] P. A. Gutierrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro and C. Hervas-Martınez, "Ordinal regression methods: survey and experimental study," IEEE Transactions on Knowledge and Data Engineering, vol. 28, pp. 127-146, 2015.

[7] P. A. Gutiérrez and S. García, "Current prospects on ordinal and monotonic classification," Progress in Artificial Intelligence, vol. 5, no. 3, pp. 171-179, 2016.

[8] D. A. Al-Qudah, A. M. Al-Zoubi, A. I. Cristea, J. J. Merelo-Guervós, P. A. Castillo and H. Faris, "Prediction of sentiment polarity in restaurant reviews using an ordinal regression approach based on evolutionary XGBoost," PeerJ Computer Science, vol. 11, p. e2370, 2025.

[9] V. M. Vargas, P. A. Gutiérrez and C. Hervás-Martínez, "Cumulative link models for deep ordinal classification," Neurocomputing, vol. 401, pp. 48-58, 2020.

[10] J. Barbero-Gomez, P.-A. Gutiérrez, V.-M. Vargas, J.-A. Vallejo-Casas and C. Hervás-Martínez, "An ordinal CNN approach for the assessment of neurological damage in parkinson's disease," Expert Systems with Applications, vol. 182, p. 115271, 2021.

[11] P. McCullagh, "Proportional-odds model," Encyclopedia of Biostatistics, vol. 6, 2005.

[12] F. García-García, D.-J. Lee, P. P. E. Yandiola, I. U. Landa, J. Martínez-Minaya, M. Hayet-Otero, M. N. Ermecheo, J. M. Quintana, R. Menéndez, A. Torres and R. Z. Jorge, "Cost-sensitive ordinal classification methods to predict SARS-CoV-2 pneumonia severity," IEEE Journal of Biomedical and Health Informatics, 2024.

[13] X. Jiang, X. Liu, Y. Wu and D. Yang, "'White Wine Quality Prediction and Analysis with Machine Learning Techniques," on Reserach Gate, 2023.

[14] E. Frank and M. Hall, "A simple approach to ordinal classification," in Machine Learning: ECML 2001: 12th European Conference on Machine Learning Freiburg, Germany, September 5--7, 2001 Proceedings 12, Springer, 2001, pp. 145-156.

[15] P. YJldJrJm, U. K. Birant and D. Birant, "EBOC: Ensemble-Based Ordinal Classification in Transportation," Journal of Advanced Transportation, vol. 2019, p. 7482138, 2019.

[16] C. KUCUK, D. BIRANT and P. Y. TASER, "A Novel Machine Learning Approach: Soil Temperature Ordinal Classification," Journal of Agricultural Sciences, vol. 28, no. 4, pp. 635--649, 2022.

[17] K. Dembczyński, W. Kotłowski and R. Słowiński, "Ordinal classification with decision rules," in Mining Complex Data: ECML/PKDD 2007 Third International Workshop, MCD 2007, Warsaw, Poland, September 17-21, 2007, Revised Selected Papers 3, Springer, 2008, pp. 169-181.

[18] J. S. Cardoso and J. F. P. d. Costa, "Learning to classify ordinal data: The data replication method," Journal of Machine Learning Research, vol. 8, no. 50, pp. 1393-1429, 2007.

[19] V. M. Vargas, A. M. Gómez-Orellana, P. A. Gutiérrez, C. Hervás-Martínez and D. Guijo-Rubio, "EBANO: A novel Ensemble BAsed on uNimodal Ordinal classifiers for the prediction of significant wave height," Knowledge-Based Systems, vol. 300, p. 112223, 2024.

[20] C. Beckham and C. Pal, "Unimodal probability distributions for deep ordinal classification," in International Conference on Machine Learning, PMLR, 2017, pp. 411-419.

[21] V. M. Vargas, A. M. Durán-Rosal, D. Guijo-Rubio, P. A. Gutiérrez and C. Hervás-Martínez, "Generalised triangular distributions for ordinal deep learning: Novel proposal and optimisation," Information Sciences, vol. 648, p. 119606, 2023.

[22] X. Liu, F. Fan, L. Kong, Z. Diao, W. Xie, J. Lu and J. You, "Unimodal regularized neuron stick-breaking for ordinal classification," Neurocomputing, vol. 388, pp. 34-44, 2020.

[23] V. M. Vargas, P. A. Gutiérrez and C. Hervás-Martínez, "Unimodal regularisation based on beta distribution for deep ordinal regression," Pattern Recognition, vol. 122, p. 108310, 2022.

[24] M. Kelly, R. Longjohn and K. Nottingham, "The UCI Machine Learning Repository," [Online]. Available: https://archive.ics.uci.edu.

[25] A. Agresti, Analysis of ordinal categorical data, John Wiley & Sons, 2010.

[26] S. Baccianella, A. Esuli and F. Sebastiani, "Evaluation measures for ordinal regression," in 2009 Ninth international conference on intelligent systems design and applications, IEEE, 2009, pp. 283-287.

[27] J. Sim and C. C. Wright, "The kappa statistic in reliability studies: use, interpretation, and sample size requirements," Physical therapy, vol. 85, no. 3, pp. 257-268, 2005.

[28] M. Cruz-Ramírez, C. Hervás-Martínez, J. Sánchez-Monedero and P. Gutiérrez, "Metrics to guide a multi-objective evolutionary algorithm," Neurocomputing, vol. 135, pp. 21-31, 2014.