International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Implementing a Hybrid Model using K-Means Clustering and Artificial Neural Networks for Risk Prediction in Life Insurance

Jeff Kimanga Nthenge[1*], Dr. Faith Mueni Musyoka[2] and Dr. David Muchangi Mugo[3]
[1]Student, Department of Computing and Information Technology, University of Embu, Embu, Kenya
e-mail: kimangajeff@gmail.com
[2]Lecturer, Department of Computing and Information Technology, University of Embu, Embu, Kenya
e-mail: mueni.faith@embuni.ac.ke
[3]Senior Lecturer, Department of Computing and Information Technology, University of Embu, Embu, Kenya
e-mail: david.mugo@embuni.ac.ke

Abstract—Accurate assessment of policy holder risk is critical for life insurance companies to properly price policies and manage long-term liabilities. However, the complexity of risk factors makes reliance solely on traditional actuarial models inadequate, especially with the proliferation of big data and unstandardized data from diverse sources. This study investigated the development and performance of a hybrid machine learning model combining artificial neural networks (ANN) and K-means clustering for enhanced risk prediction in life insurance underwriting. The exponential growth of unlabelled data presented challenges for predictive modelling. The proposed hybrid model leveraged the strengths of artificial neural networks in modelling nonlinear relationships and K-means clustering in unsupervised for pattern recognition to handle unstandardized data. Using anonymized life insurance application data from Kaggle, the hybrid model was evaluated against the artificial neural network algorithm alone. The results demonstrated that integrating K-means clustering and artificial neural network together with principal component analysis for pre-processing led to superior model performance, with testing accuracy improving from 90% for artificial neural network to 98% for the hybrid technique. Additional metrics like precision, recall and AUC also showed enhancements. The improved predictive capability highlighted the potential of the hybrid approach in transforming legacy underwriting practices towards a more sophisticated data-driven analytical evaluation of policy holder risk. However, limitations existed including the use of single sourced insurance dataset due to data privacy concerns. Further research on integrating diverse algorithms can help insurers unlock more value and gain a competitive edge through advanced analytical modelling and testing on larger real-world datasets. While challenges remain, this study provided key insights into a promising new technique for modernizing risk prediction in the life insurance industry in the era of big data.

Keywords- Artificial Neural Networks, ANN, K-Means Clustering, Hybrid Model, PCA, Risk Prediction, Life Insurance.

I. INTRODUCTION

Life insurance underwriting involves evaluating numerous risk factors to determine premiums and make policy acceptance decisions [1]. Underwriters traditionally used predetermined rules and legacy actuarial models which faced challenges in modelling nonlinear interactions, missing data, and capturing complex relationships between diverse risk variables [2]. With the exponential growth of unstructured big data sources and machine learning advancements, data-driven predictive analytics has emerged as a powerful tool to modernize risk modelling in insurance underwriting [3]. However, unstandardized data from disparate sources and variety in structured and unstructured data formats present new complexities.

Artificial Neural Networks (ANN) have shown effectiveness in modelling nonlinear patterns between applicant attributes like age, lifestyle, medical history, and risk level outcomes [4]. However, ANN models face limitations like overfitting, sensitivity to hyperparameters and lack of transparency [5]. On the other hand, unsupervised clustering techniques like K-means can derive insights from unlabeled data, but optimizing clusters remains challenging [6]. Hybrid machine learning models that integrate supervised and unsupervised algorithms have demonstrated better performance compared to individual algorithm techniques [7]. However, applications tailored to life insurance risk prediction remain relatively unexplored. This highlighted the need and motivation for developing a hybrid model by combining Artificial Neural Network and K-means Clustering to cater to the unique challenges of life insurance data complexity, variety, and model performance requirements.

The hybrid approach leveraged ANN's nonlinear modelling capabilities while K-means unsupervised feature learning and pattern recognition were used to achieve more accurate analytical policy holder risk evaluation. The objective was to develop and empirically evaluate a hybrid ANN and K-means model using life insurance dataset, and assess performance based on metrics including accuracy, precision, recall, and AUC. The

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

research contributions included a tailored predictive modelling technique for enhancing actuarial risk models using integrated machine learning. It provided model implementation guidance focused on transparent and ethical AI. However, limitations persisted including the use of secondary datasets. Privacy and omitted risk factors also remain concerns. Notwithstanding these constraints, the study represented an important step towards modernizing legacy underwriting practices through data-driven analytical sophistication. It sets the direction for industry-academia collaboration in developing robust hybrid machine learning models that meet regulatory rigor while delivering superior analytical value.

II. LITERATURE REVIEW

Accurate evaluation of risk is fundamental to pricing and profitability across the insurance industry. Actuarial models using statistical techniques have traditionally been used to estimate risk. However, these face limitations in capturing the complexity arising from the exponential growth of unstandardized big data from multiple sources and inherent nonlinear relationships between diverse risk parameters [8]. With advancements in computational capabilities and machine learning algorithms, data-driven predictive modelling has emerged as a promising technology to enhance analytical risk evaluation [9]. However, supervised learning models designed for structured data and unsupervised algorithms for unstructured, individually have not fully addressed the challenges presented by unstructured data's variety, veracity, and velocity.

A. Artificial Neural Networks for Risk Prediction

Artificial neural networks (ANN) have shown effectiveness in modelling nonlinear patterns between diverse variables related to demographics, behaviors, environmental factors and predicted outcomes [10]. Their interconnected neuron layers can detect complex relationships, making ANN well-suited for combining multidimensional risk factors from intricate insurance datasets. For instance, [4] a study developed an ANN model with over 100 input variables, demonstrating high accuracy in categorizing general risk levels. The study presented a pure ANN architecture with an average correct classification of 98.26% for financial risk classification. Such results highlight ANN's potential for accurate prediction given substantial training data representing populations with adequate diversity. However, ANN remains prone to overfitting, especially with limited heterogeneous datasets [11]. Research shows that ANN can outperform traditional statistical techniques in identifying predictive combinations from multifaceted data [12]. However, ensemble machine learning algorithms like random forests and REPT trees have also shown superior accuracy to ANN in some studies, underlining the need for hyperparameter optimization [13–15]. Overall, ANN provides more sophisticated risk modelling than legacy statistical models, but ensemble methods may offer better performance.

B. K-Means Clustering for handling Data Complexity

The unsupervised K-means algorithm clusters similar data points, revealing insights from unlabeled data [6,16]. Studies have shown that integrating K-means with ANN can improve prediction accuracy by pre-processing data through dimensionality reduction and pattern detection [17–21]. Analyzing new data with initially unknown properties benefits from K-means identifying behaviors and attributes associated with each cluster. This uncovers hidden relationships in unstructured and unlabeled raw data that can inform predictive modelling [21]. Clustering -driven data segmentation also helps mitigate overfitting risks [22]. However, challenges remain in effectively clustering unstructured data with outliers and determining optimal clusters [23]. This underscores the need for hybrid approaches that leverage K-means while addressing its limitations.

C. Hybrid Machine Learning Models

Given their complementary capabilities, integrated hybrid machine learning models combining ANN and K-means clustering algorithms show strong promise for enhancing predictive accuracy by handling data complexity effectively. A study by Kaya et al. 2018, improved accuracy by using K-means clustering for feature selection with ANN [19]. While another study found incorporating clustering pre-processing led to increased model sensitivity [24]. Biswas et al. 2021 proposed a hybrid architecture applying K-means segmentation followed by ANN-based classification, benefiting from both unsupervised grouping and supervised learning [17]. Such studies highlight that thoughtfully engineered hybrid pipelines that harness K-means unsupervised feature learning and ANN's predictive prowess could potentially enhance analytical evaluation over individual techniques. However, research focused on tightly integrated hybrid architectures tailored to address data intricacies remains relatively scarce, highlighting a significant literature gap. While ANN and K-means have shown individual effectiveness for risk modelling, a hybrid approach that closely coordinates the two techniques exhibits greater potential based on their complementary strengths. Notably, substantial research opportunities remain to rigorously develop tailored holistic hybrid architectures suited to the challenges of multidimensional, unstructured data and precise predictive requirements. Meticulously designed integrated models could significantly advance analytical risk evaluation capabilities.

III. METHODS

A. Research Design

A quasi-experimental design for research was adopted since real underwriting data could not be accessed given confidentiality reasons [25]. The hybrid ANN and K-means model was compared to the standalone ANN model to assess performance improvement. Fig. 1 below shows the high-level approach followed for the development of the model.

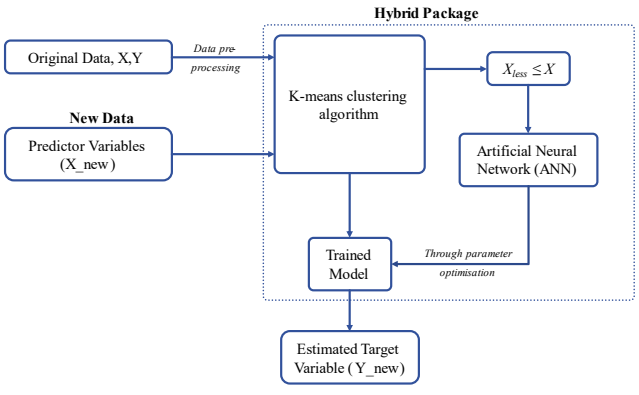

Figure 1. Hybrid model framework

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Publicly available anonymized life insurance data from Kaggle was used. The independent variables were applicant attributes like age, height, medical history etc. The dependent variable was risk rating for the policy holder.

B. Data collection and Pre-processing

Anonymized life insurance application data for 59,381 applicants was obtained from a Kaggle repository (Kaggle, 2017). Data pre-processing involved handling missing values, one-hot encoding categorical variables, and min-max feature scaling to ready the dataset for machine learning among others [26].

C. Tools and Environment

The Python programming language was used given the flexibility, scalability, and vast machine learning libraries like Scikit-Learn, Keras and TensorFlow [27]. The Google Colab platform was used for developing and executing the models using GPU acceleration [28].

D. Hybrid Model Development

The hybrid model was developed by integrating ANN and K-means clustering algorithms. K-means clustering was first applied to a group of similar applicants and reduced the dimensionality of the dataset. The optimal number of clusters was determined using the elbow method [29]. The clustered data was input to the ANN model for policy holder risk classification. The ANN model was then trained using backpropagation and stochastic gradient descent optimization [18]. In addition, the number of hidden layers, epochs and batch size were tuned to find the optimal architecture.

E. Model Evaluation

The hybrid model was evaluated using accuracy, precision, recall and f1 score metrics on the test dataset and compared to the standalone ANN model [30,31]. Moreover, receiver operating characteristic (ROC) curve analysis was conducted to assess predictive capability and the Area Under the Curve (AUC) metric was used to evaluate model discrimination ability [32].

IV. RESULTS

This study aimed to develop and evaluate a hybrid machine learning model combining artificial neural networks (ANN) and K-means clustering to improve analytical risk prediction demonstrated through policy holder risk modelling in life insurance underwriting. This section presents the key findings of model development, and evaluation to assess the hybrid model's performance against ANN alone.

A. Gaps in Artificial Neural Networks and K-Means Clustering for Risk Prediction

A systematic literature review was conducted to identify limitations and gaps in using ANN and K-means Clustering for insurance risk prediction based on previous studies. Table 1 summarizes the key gaps identified from the literature review. These gaps highlighted areas for improvement that were addressed in developing the hybrid model in this study through comprehensive tuning, testing and evaluation of the hybrid model.

TABLE I. SUMMARY OF GAPS IN A. ARTIFICIAL NEURAL NETWORKS AND K-MEANS CLUSTERING FROM LITERATURE

Study	Gap
[23]	Difficulty in determining optimal clusters in K-means
[20]	Limited evaluation of hybrid models with other techniques like feature selection
[33]	Limited testing of K-means hybrid model in real scenarios
[34]	Lack of large datasets and tuning options for the ANN model
[35]	Limited model evaluation metrics and dataset diversity for ANN
[36]	Lack of comprehensive metrics to evaluate ANN performance
[37]	Lack of comparison of the ANN model to other risk models

B. Development of the Hybrid Model

Guided by the gaps identified, a systematic process was followed to develop the hybrid ANN and K-means model for risk prediction. The anonymized life insurance dataset from the Kaggle repository included 59,381 applicants and 128 features. The features included demographics, medical history and insurance details as highlighted in Table 2 below:

Data pre-processing was performed to prepare the features for machine learning compatibility. Several techniques were applied to transform the data, including missing value imputation through mean substitution. Categorical variables were also handled during the pre-processing stage. One-hot encoding was applied to convert categorical variables into a binary representation, creating separate columns for each unique category. This process enabled the machine learning algorithms to effectively interpret and utilize the categorical data. Additionally, feature scaling was employed to normalize the range of the numerical features. The min-max normalization technique was utilized, which rescaled the feature values to a specific range (typically between 0 and 1). This normalization process ensured that all features were on a similar scale, preventing any feature from dominating the model's learning process.

TABLE II. SAMPLE FEATURES IN DATASET

Variable	Description
Id	A unique identifier associated with an application.
Product_Info_1-7	A set of normalized variables relating to the product applied for
Ins_Age	Normalized age of the applicant
Ht	Normalized height of the applicant
Wt	Normalized weight of the applicant
BMI	Normalized BMI of the applicant
Employment_Info_1-6	A set of normalized variables relating to the employment history of the applicant.
InsuredInfo_1-6	A set of normalized variables providing information about the applicant
Insurance_History_1-9	A set of normalized variables relating to the insurance history of the applicant.
Family_Hist_1-5	A set of normalized variables relating to the family history of the applicant.
Medical_History_1-41	A set of normalized variables relating to the medical history of the applicant.
Medical_Keyword_1-48	A set of dummy variables relating to the presence/absence of a medical keyword being associated with the application.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Variable	Description
Response	This is the target variable, an ordinal variable relating to the final decision associated with an application

C. Exploratory Data Analysis

Through the process of exploratory analysis, significant data characteristics were identified. There existed a highly imbalanced distribution in the risk variable, with a clustering of observations in the higher risk categories, as can be inferred from Fig. 2 Secondly, examination of the data revealed the existence of outliers in certain features such as employment history, as shown in Fig. 3. Finally, a high degree of correlation between features, specifically BMI and weight, was observed, as illustrated in Fig. 4. These insights guided data pre-processing and feature engineering decisions.

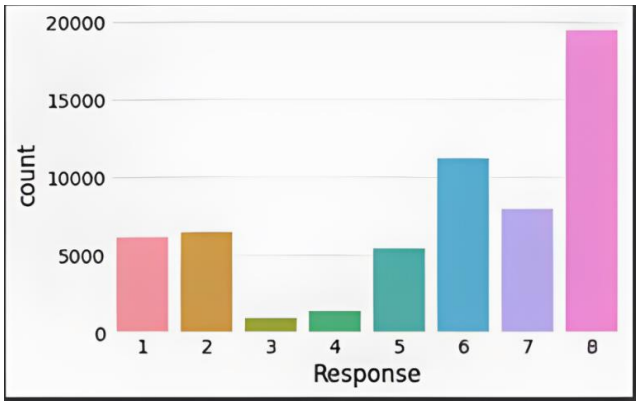

Figure 2. Imbalanced risk variable distribution

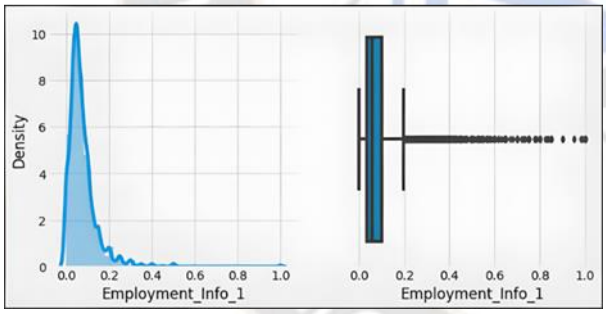

Figure 3. Outliers in employment feature

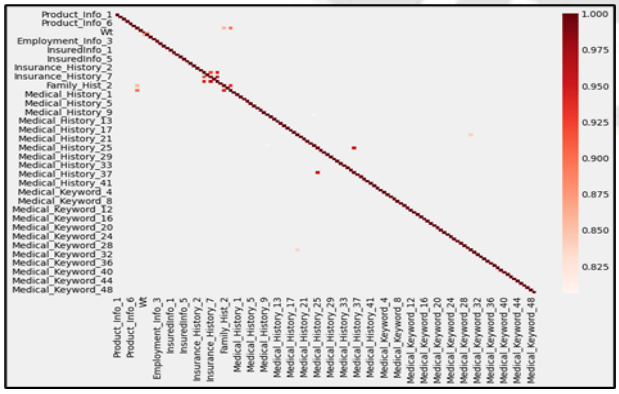

Figure 4. Correlation matrix showing highly correlated features

D. Dimensionality Reduction using Principal Component Analysis

Principal component analysis (PCA) was applied to understand the most significant sources of variation within the dataset. This was meant to help in understanding whether there are particularly useful features in the dataset that should be preserved during feature selection. As shown in the table below, the variables that appeared most frequently were medical_history_4 and medical_history_41. It was therefore more important for the models to place greater importance on these features later.

TABLE III. FREQUENCY OF SIGNIFICANT VARIABLES BASED ON PRINCIPAL COMPONENT ANALYSIS

Variable	Principal Component	PC Count
Medical_History_4	PC3,PC5,PC7,PC9,PC10	5
Medical_History_41	PC5,PC7,PC9,PC10,PC11	5
Medical_History_16	PC10,PC11,PC15,PC16	4
Product_Info_4	PC25,PC27,PC28,PC30	4
Medical_Keyword_22	PC35,PC37,PC38,PC40	4
Medical_History_2	PC23,PC24,PC25,PC26	4
Product_Info_2_D1	PC15,PC17,PC18,PC26	4
Employment_Info_3	PC10,PC12,PC15,PC23	4
Medical_Keyword_11	PC29,PC30,PC31	3
Employment_Info_5	PC12,PC15,PC23	3
Medical_History_34	PC14,PC15,PC16	3
Family_Hist_1	PC32,PC33,PC34	3
Medical_History_13	PC12,PC13,PC14	3
Product_Info_6	PC9,PC12,PC13	3

E. ANN Model Evaluation

Before developing the hybrid model, the ANN algorithm was independently evaluated on the dataset. This formed the baseline for comparative assessment. The ANN model was trained for 20 epochs using stochastic gradient descent optimization and binary cross-entropy loss function. Training performance improved progressively across epochs as evidenced by reducing loss and improving accuracy as shown in Table 4.

TABLE IV. TRAINING ACCURACY AND LOSS CURVE FOR A. ARTIFICIAL NEURAL NETWORK MODEL

Epoch	Time	Loss	Accuracy	Validation Loss	Validation Accuracy
2	5s	1.382	0.7567	1.1569	0.7859
3	9s	1.0362	0.8035	0.8984	0.8301
4	7s	0.8382	0.8292	0.7479	0.8439
5	9s	0.7188	0.8439	0.6544	0.855
6	8s	0.6405	0.8533	0.5892	0.8655
7	9s	0.5857	0.8601	0.5437	0.8701
8	5s	0.5453	0.8667	0.5085	0.8746
9	6s	0.5142	0.8712	0.4817	0.8774

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Epoch	Time	Loss	Accuracy	Validation Loss	Validation Accuracy
10	4s	0.4896	0.8746	0.4599	0.8835
11	4s	0.4696	0.8784	0.4426	0.8856
12	5s	0.4531	0.8807	0.4266	0.8889
13	5s	0.4391	0.8835	0.4146	0.89
14	4s	0.4271	0.8854	0.4036	0.8923
15	4s	0.4169	0.8871	0.3938	0.8934
16	5s	0.4077	0.8888	0.3863	0.8941
17	4s	0.3998	0.8905	0.3785	0.8949
18	4s	0.3928	0.8916	0.3719	0.8959
19	6s	0.3863	0.8924	0.3664	0.897
20	4s	0.3807	0.8935	0.3614	0.8992

On the unseen test set, the ANN model achieved an accuracy of 90% and an AUC of 0.95 as highlighted in Table 5 below. The precision, recall and F1-score were also reasonably high indicating good predictive performance. This analysis established baseline ANN performance to benchmark the hybrid model against.

TABLE V. SUMMARY OF GAPS IN ANN AND K-MEANS FROM LITERATURE

Metric	Value
Accuracy	0.9
AUC	0.95
Precision	0.9
Recall	0.9
F1-score	0.9

F. Hybrid Model Architecture

A systematic approach was used to develop a hybrid architecture utilizing K-means Clustering and an Artificial Neural Network (ANN). Specifically, the pre-processed dataset of applicants was subjected to K-means clustering to both groups of similar applicants and reduced dimensionality. The number of optimal clusters was identified as 15 using the elbow method, as depicted in Fig. 5, resulting in well-separated compact clusters. Reducing K-means clusters from 15 to 10 did not impact accuracy but increased model training time slightly by 1 hour indicating that clustering parameters require thorough tuning.

The clustered dataset was utilized as input to the ANN model for risk classification. The hyperparameters of the ANN model, including the hidden layers, epochs, and batch size, were carefully selected using grid search to ensure optimal performance. The proposed hybrid architecture effectively leveraged the strengths of K-means Clustering and ANN for accurate and efficient classification of policy holder risk. This was achieved through the creation of well-defined, compact clusters via K-means clustering and the use of ANN for classification.

The results showed the potential of effectively analyzing and processing large datasets using a hybrid approach. In this experiment, a clustering algorithm was utilized to identify the optimal number of clusters for the dataset under examination.

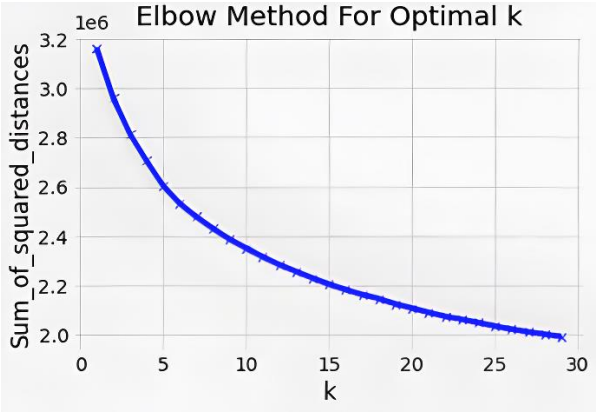

Figure 5. Elbow plot to determine optimal number of clusters

Through elbow method, it was found to be 15, as it resulted in the lowest sum of squared distances of 2.2 which indicated the similarity within a cluster [29]. The results further emphasized the effectiveness of clustering unlabeled insurance data into significant segments that helped to balance segmentation granularity and model complexity. Moreover, it should be noted that while K-means Clustering demonstrated suitable groupings for the complex insurance dataset, the technique has certain limitations such as sensitivity to outliers, which necessitated a hybrid approach. The successful generation of suitable homogeneous groupings using K-means provided valuable input for the ANN model. These findings revealed the potential of utilizing clustering techniques for unsupervised learning in generating meaningful segments in complex datasets.

G. Hybrid Model Evaluation

The hybrid K-means + ANN model was evaluated on the test set across key metrics and compared to the ANN baselines. Training progress was monitored across epochs for the hybrid model as shown in Table 6. A similar decreasing loss reading and increasing accuracy readings was observed indicating progressive optimization.

TABLE VI. ACCURACY AND LOSS DATA DURING HYBRID MODEL TRAINING

Epoch	Time	Loss		Accuracy	
		ANN	*ANN+ K-means*	*ANN*	*ANN+ K-means*
2	5s	1.382	1.1569	0.7567	0.7859
3	9s	1.0362	0.8984	0.8035	0.8301
4	7s	0.8382	0.7479	0.8292	0.8439
5	9s	0.7188	0.6544	0.8439	0.855
6	8s	0.6405	0.5892	0.8533	0.8655
7	9s	0.5857	0.5437	0.8601	0.8701
8	5s	0.5453	0.5085	0.8667	0.8746
9	6s	0.5142	0.4817	0.8712	0.8774
10	4s	0.4896	0.4599	0.8746	0.8835
11	4s	0.4696	0.4426	0.8784	0.8856

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Epoch	Time	Loss		Accuracy	
		ANN	*ANN+ K-means*	*ANN*	*ANN+ K-means*
12	5s	0.4531	0.4266	0.8807	0.8889
13	5s	0.4391	0.4146	0.8835	0.89
14	4s	0.4271	0.4036	0.8854	0.8923
15	4s	0.4169	0.3938	0.8871	0.8934
16	5s	0.4077	0.3863	0.8888	0.8941
17	4s	0.3998	0.3785	0.8905	0.8949
18	4s	0.3928	0.3719	0.8916	0.8959
19	6s	0.3863	0.3664	0.8924	0.897
20	4s	0.3807	0.3614	0.8935	0.98

1) Accuracy, AUC, and related metrics

The hybrid model achieved a test accuracy of 98% compared to 90% for ANN reflecting the positive impact of clustering. AUC improved from 0.95 to 0.98 with the hybrid model highlighting better separation. Precision, recall and F1-score also showed improvements as evidenced in Table 7. The hybrid model demonstrated superior performance over ANN across key metrics indicating the benefits of integrated clustering.

TABLE VII. COMPARISON OF HYBRID AND ARTIFICIAL NEURAL NETWORK ON TEST SET

Metric	ANN	Hybrid	Improvement
Accuracy	0.9	0.98	9%
AUC	0.95	0.98	3%
Precision	0.9	0.98	8%
Recall	0.9	0.98	8%
F1-score	0.9	0.98	8%

2) Validation using Logistic Regression

As seen in Table 8 below the regression analysis revealed an intercept of 0.9 and a coefficient of 0.08 for the Hybrid Model. The R2 score was found to be 1.0, indicating that the model accounted for all the variances seen in the data. Overall, the analysis results implied that the Hybrid Model had a better accuracy than the ANN model.

TABLE VIII. COMPARISON OF HYBRID AND ARTIFICIAL NEURAL NETWORKS ON TEST SET

Metric	Value
Intercept	0.9
Coefficient	0.08
R-squared score	1

To further analyze and validate the performance of the ANN and hybrid model, Logistic Regression on the same dataset was used for training and testing the models while experimenting with ANN and the hybrid model to get the mean squared error (MSE) and R2 score and as seen in Table 9 below, Hybrid model had a smaller MSE of 69.978813 compared to the ANN model's MSE of 117.242058. These results indicated that the hybrid model's MSE and R2 score outperforms the ANN model.

TABLE IX. LOGISTIC REGRESSION ON DATASET USED FOR TRAINING AND TESTING THE MODELS

Dummy Variable	Model	MSE	R^2 score
0	ANN	117.242058	-0.598746
1	Hybrid	69.978813	0.045749

The scatter plot also showed clear improvements with the hybrid model as visualized in Fig. 6.

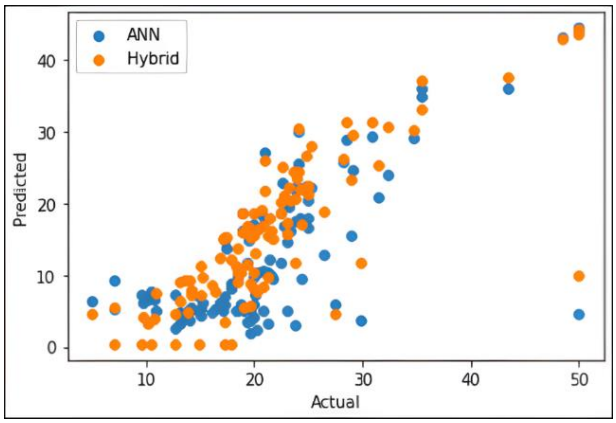

Figure 6. Elbow plot to determine optimal number of clusters

3) Optimizer Performance

Different optimizers were used in the hybrid model and evaluated. The Adam optimizer yielded the highest test accuracy of 97% as shown in Table 10. This aligned with findings from previous research that the Adam optimizer provides faster convergence for ANN models.

TABLE X. PERFORMANCE COMPARISON OF DIFFERENT OPTIMIZERS FOR HYBRID MODEL

Optimizer	Mean Accuracy	Standard Deviation
SGD	0.888983	0.005568
RMSprop	0.975617	0.002475
AdaGrad	0.837333	0.008196
Adadelta	0.71395	0.014775
Adam	0.975883	0.001899
AdaMax	0.963483	0.002736
Nadam	0.97645	0.001996

V. DISCUSSION

This study presented a systematic process for developing and evaluating a hybrid machine learning model combining artificial neural networks (ANN) and K-means clustering to improve risk prediction in life insurance underwriting. The results exhibited remarkable effectiveness of the proposed hybrid approach in enhancing the predictive accuracy of the target (Dependents Variable) from the set of independent variables fed to the system.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Testing accuracy improved substantially from 90% with standalone ANN modelling to 98% when integrating the complementary strengths of ANN and K- Means Clustering algorithms. A key factor driving these significant gains was the ability of the unsupervised K-means Clustering technique in effectively handling the challenges presented by the unstructured secondary insurance dataset.

The exploratory analyses revealed crucial data complexities including missing values, outliers, class imbalance, redundant features, and highly correlated variables. K-means helped address these intricacies through imputation, segmentation, dimensionality reduction, and pattern detection. This data pre-processing enhanced the ability of the subsequent ANN model to learn nonlinear relationships between diverse risk factors by reducing noise and redundancy. In addition, determining the optimal number of clusters ensured appropriate granularity for the ANN model without overcomplicating the architecture.

The hybrid model outperformed ANN across other key evaluation metrics like precision, recall, F1-score, and AUC. This further validated the integrated approach's effectiveness in leveraging unsupervised learning to prepare complex data for tailored ANN modelling. Hyperparameter tuning and techniques like regularization further optimized the hybrid architecture by preventing overfitting. These results empirically demonstrated the considerable benefits of a tailored machine learning pipeline that can harnesses K-means' unsupervised data handling strengths and ANN's predictive prowess for substantially improved analytical policy holder risk evaluation.

The significantly improved hybrid model predictive performance carries major implications for life insurers given the pivotal role accurate risk prediction plays in managing pricing, profitability, and core operations. The considerable gains in accuracy, precision and sensitivity have a profound impact considering the complexities of insurance data. By leveraging K-means' unsupervised learning strengths, the hybrid approach unlocked the ability to uncover predictive insights from unstructured raw data with more granularity. This facilitated more optimized pricing aligned to expected liabilities and claims. The enhanced analytical sophistication also creates opportunities to develop customized underwriting and product development strategies tailored to market niches based on refined risk segmentation.

A. Limitations

Considering the research results, several limitations must be considered when applying hybrid machine learning for risk prediction in life insurance. One limitation is the use of publicly available data, which may not capture real-world diversity compared to proprietary underwriting datasets. This constrains model accuracy and applicability. Further, limited data volume poses challenges in exploiting the full capabilities of artificial neural networks which depend on substantial training data. Small datasets increase the risk of overfitting. Data privacy regulations also pose practical constraints on accessing real underwriting data to test models, limiting viability assessments. There are ethical risks of patient data use without consent. The lack of implementation guidelines for integrating predictive models into existing underwriting workflows presents adoption barriers for insurers. Change management concerns must be addressed. It is important to note the need for ongoing monitoring frameworks to ensure models maintain prediction fidelity over time and do not inadvertently introduce bias against protected groups as data evolves.

Lastly, we have machine learning bias or AI bias which limits development and deployment of AI based systems. Some ways that bias can be found in algorithms include training data, human creators, and decision-making processes that are biased. While these limitations exist, the study still managed to provide valuable methodological insights.

VI. CONCLUSION AND FUTURE WORK

Accurate evaluation of policy holder risk is imperative for life insurers given long-term liabilities. However, the complexity of risk factors makes reliance solely on traditional actuarial techniques inadequate. This study demonstrated the significant improvements in prediction accuracy achievable by life insurers through a hybrid machine learning model combining ANN and K-means clustering. The proposed integrated approach leverages complementary strengths of the algorithms which contributed to superior accuracy, precision, and recall. The empirically validated effectiveness of the hybrid technique highlights its potential to transform legacy underwriting practices towards data-driven analytical sophistication.

Areas of future research include testing on larger real-world datasets, adding contextual policyholder data, continuously refining model performance post-implementation, and developing risk management frameworks to integrate models seamlessly into underwriting processes. The researchers can also explore underlying factors such as social media rankings, driving offences and revenue returns that can drive the risk levels of the applicants. Additionally, researchers should focus on data anonymization, synthetic dataset generation, incremental integration, and trustworthy AI practices that can help address the key challenges. Further, research can be done to tackle the risk for bias in machine learning algorithms by increasing transparency, instituting regulations for independent algorithmic audit tests, and proposing frameworks to assess evolving risks while mitigating bias.

With rigorous research-led adoption, hybrid modelling promises to redefine underwriting standards by enabling life insurers to achieve transformation in analytical capabilities, risk management outcomes, competitive positioning, and long-term profitability.

REFERENCES

[1] Dwivedi SK, Mishra A, Kumar Gupta R. Risk prediction assessment in life insurance company through dimensionality. International Journal of Scientific & Technology Research 2020;9:1528–32.

[2] Li Q, Duong TD, Wang Z, Liu S, Wang D, Xu G. Causal- Aware generative imputation for automated underwriting. Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Association for Computing Machinery; 2021, p. 3916–24.

[3] Verma A, Taneja A, Arora A. Fraud detection and frequent pattern matching in insurance claims using data mining techniques. 2017 10th International Conference on Contemporary Computing, IC3 2017, Institute of Electrical and Electronics Engineers Inc.; 2017, p. 1–7.

[4] Samuel OW, Asogbon GM, Sangaiah AK, Fang P, Li G. An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. Expert Syst Appl 2017;68:163–72.

[5] Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. J Big Data 2015;2:1–21.

[6] Uzila A. K-means Clustering and Principal Component Analysis in 10 minutes. Towards Data Science 2022. https://towardsdatascience.com/k-means-clustering-and-principal-component-analysis-in-10-minutes-2c5b69c36b6b (accessed September 30, 2022).

[7] Pes B. Ensemble feature selection for high-dimensional data: A stability analysis across multiple domains. Neural Comput Appl 2020;32:5951–73.

[8] Blier-Wong C, Cossette H, Lamontagne L, Marceau E. Machine learning in P&C insurance: A review for pricing and reserving. Risks 2021;9:1–26.

[9] Cam H. Cyber risk and vulnerability estimation. The Journal of Defense Modeling and Simulation (JDMS) 2022;19:3–4.

[10] Aziz MN. A Review on Artificial Neural Networks and its' applicability. Bangladesh Journal of Multidisciplinary Scientific Research 2020;2:48–51.

[11] Nicolae-Eugen C. Lowering evolved Artificial Neural Network overfitting through high-probability mutation. 2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Institute of Electrical and Electronics Engineers Inc.; 2016, p. 325–9.

[12] Manni A, Saviano G, Bonelli MG. Optimization of the ANNs predictive capability using the taguchi approach: A case study. Mathematics 2021;9:1–16.

[13] Boodhun N, Jayabalan M. Risk prediction in life insurance industry using supervised learning algorithms. Complex & Intelligent Systems 2018;4:145–54.

[14] Jain R, Alzubi JA, Jain N, Joshi P. Assessing risk in life insurance using ensemble learning. Journal of Intelligent & Fuzzy Systems 2019;37:2969–80.

[15] Kamil A, Hassan I, Abraham A. Modeling insurance fraud detection using ensemble combining classification. vol. 8. 2016.

[16] Wu T, Xiao Y, Guo M, Nie F. A general framework for dimensionality reduction of K-Means Clustering. J Classif 2020;37:616–31.

[17] Biswas A, Islam MS. Brain tumor types classification using K-means Clustering and ANN approach. International Conference on Robotics, Electrical and Signal Processing Techniques, 2021, p. 654–8.

[18] Atil H, Akilli A. Comparison of Artificial Neural Network and K-means for clustering dairy cattle. International Journal of Sustainable Agricultural Management and Informatics 2016;2:40–52.

[19] Kaya U, Yılmaz A, Şaykol E. Designing a neural network model using K-means Clustering for risk analysis of lung cancer disease. Journal of Aeronautics and Space Technologies 2018;11:107–18.

[20] Malav A, Kadam K, Kamat P. Prediction of heart disease using K-means and Artificial Neural Network as hybrid approach to improve accuracy. International Journal of Engineering and Technology 2017;9:3081–5.

[21] Hartono, Sitompul OS, Tulus, Nababan EB. Optimization Model of K-Means Clustering Using Artificial Neural Networks to handle class imbalance problem. IOP Conf Ser Mater Sci Eng 2018:1–9.

[22] Nidheesh N, Abdul Nazeer KA, Ameer PM. An enhanced deterministic K-Means Clustering algorithm for cancer subtype prediction from gene expression data. Comput Biol Med 2017;91:213–21.

[23] Orong MY, Sison AM, Medina RP. A hybrid prediction model integrating a modified genetic algorithm to K-means segmentation and C4.5. TENCON 2018-2018 IEEE Region 10 Conference, IEEE; 2019, p. 1853–8.

[24] Zhu C, Idemudia CU, Feng W. Improved Logistic Regression model for diabetes prediction by integrating PCA and K-means techniques. Inform Med Unlocked 2019;17:1–7.

[25] Bunselmeyer E, Schulz P. Quasi-experimental research designs as a tool for assessing the impact of transitional justice instruments. The International Journal of Human Rights 2019;24:688–709.

[26] Anunaya S. Data preprocessing in data mining - A hands on guide. Analytics Vidhya 2021. https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/ (accessed July 1, 2022).

[27] Sharma A, Khan F, Sharma D, Gupta S. Python: The programming language of future. International Journal of Innovative Research in Technology 2020;6:115–8.

[28] Gunawan TS, Ashraf A, Riza BS, Haryanto EV, Rosnelly R, Kartiwi M, et al. Development of video-based emotion recognition using deep learning with Google Colab. Telkomnika (Telecommunication Computing Electronics and Control) 2020;18:2463–71.

[29] Shi C, Wei B, Wei S, Wang W, Liu H, Liu J. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP J Wirel Commun Netw 2021;2021:1–16.

[30] Dinga R, Penninx BWJH, Veltman DJ, Schmaal L, Marquand AF. Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. BioRxiv 2019:743138.

[31] Arefin MS, Kaiser MS, Bandyopadhyay A, Ahad AR, Ray K. Proceedings of the international conference on Big Data, IoT, and Machine Learning. Springer Singapore, 2022.

[32] Guo X, Sun Z, Jiang S, Jin X, Wang H. Identification and validation of a two-gene metabolic signature for survival prediction in patients with kidney renal clear cell carcinoma. Aging 2021;13:8276–89.

[33] Verma V, Bhardwaj S, Singh H. A hybrid K-mean Clustering algorithm for prediction analysis. Indian J Sci Technol 2016;9:1–5.

[34] Pal R, Sekh AA, Kar S, Prasad DK. Neural network based country wise risk prediction of COVID-19. Applied Sciences 2020;10:1–16.

[35] Yang D. Evaluation of enterprise financial risk level under digital transformation with Artificial Neural Network. Security and Communication Networks 2022;2022:1–9.

[36] Kumar Gupta D, Goyal S. Credit risk prediction using Artificial Neural Network algorithm. International Journal of Modern Education and Computer Science 2018;10:9–16.

[37] Radosteva M, Soloviev V, Ivanyuk V, Tsvirkun A. Use of neural network models in the market risk management. Adv Syst Sci Appl 2018;18:53–8.