Prediction of System Usability through Neural NetworksApplied to Conversational User Interfaces

Lady K. Gomez
Autonomous University of Zacatecas,

Zacatecas, Mexico, 98060,
ladykatherine@unicauca.edu.co

and

Huizilopoztli Luna-Garcia
Autonomous University of Zacatecas,
Zacatecas, Mexico, 98060,
hlugar@uaz.edu

and

Cesar A. Collazos
University of Cauca,
Cauca, Colombia, 19004,
ccollazo@unicauca.edu.co

Abstract

This study aims to explore the potential of Deep Neural Networks (DNNs) to predict perceived usability in Conversational User Interfaces (CUIs). The research focuses on two systems, ChatGPT and Gemini, evaluated in two interaction modalities: text and voice. The main objective is to show that models based on Deep Neural Networks (DNN) can be used as predictive tools to estimate perceived usability. To this end, a structured methodology is applied, including data segmentation, integration of demographic variables, regularization techniques, and cross validation. Through this approach, we hope to determine how demographic and experience factors influence usability prediction and whether voice interaction provides a more stable framework for its evaluation. The expected outcome is the development of a predictive model capable of supporting usercentered design decisions and contributing to the optimization and personalization of experiences with conversational systems.

Keywords: CUI, Perceived Usability, Deep Neural Networks, SUS, VUS.

1 Introduction

In recent years, various voice systems or conversational user interfaces (CUIs) have gained importance in many areas, ranging from education to entertainment [1, 2]. The growth of these systems demonstrates the importance of focusing not only on their functionality but also on the quality of the user experience, with an emphasis on usability [3].

This study is presented as an extended work of the article "Review of Heuristics and Usability and UX Scales for Speech-Based Technologies" [4], in which a systematic review was conducted on scales, heuristics, and design patterns in speech-based technological systems such as CUIs, VUIs, among others. That work outlined future research aimed at a more focused study of conversational user interfaces (CUIs) and identified a research gap in the validation and creation of scales, heuristics, and design guidelines [4].

To address this gap, a neural network architecture is proposed to predict usability scores using the SUS [5] and VUS [6] scales for CUI systems, in order to estimate the usability perceived by users in both voice and text modalities. For this purpose, a strategic methodological approach was adopted to understand data segmentation and identify key patterns in the experiences of a group of users. In addition, a comparison is included between the predictions generated by the models and the actual evaluations provided by the users, representing a significant advancement over the previous work. The study is based on the premise that the interaction method, together with the conversational system, combined with demographic factors and prior experiences, directly influences users' perception of usability. For this purpose, various models were evaluated for each combination of system and modality, with the aim of comparing their performance and determining under which conditions a more accurate prediction can be achieved.

The results obtained not only corroborate the feasibility of implementing deep learning models in this field, but also provide an empirical foundation for progressing toward a more personalized and adaptable design of conversational systems, in line with users' genuine expectations and requirements. User-Centered Design (UCD) [7] focuses on involving users in the creation of systems; however, these processes can become costly, time consuming, difficult to apply iteratively within design processes, and require a considerable number of users for the information to be sufficient to capture users' perceptions of usability. Considering this context, predictive models do not seek to replace user based evaluation, but rather to complement it by reducing the number of users needed, while enabling the anticipation of problems and supporting decision-making during the design process. Systems developed using UCD are characterized by being evaluated iteratively and continuously [7]; among the most common approaches are usability evaluation methods, which primarily assess the efficiency, effectiveness, and satisfaction [5] of the final system in order to make adjustments to the system.

2 Related Works

Several studies have explored the use of machine learning techniques in usability and user experience assessment. Previous research has applied models such as Support Vector Machines (SVM), decision trees, and artificial neural networks to predict usability or satisfaction metrics in educational systems, mobile applications, and web platforms. These studies demonstrate the feasibility of the approach, although most are limited to non-conversational contexts or small, specific datasets, which restricts the generalizability of their results [8]. Regarding the evaluation of conversational interfaces, recent literature has proposed specific scales such as the Voice Usability Scale (VUS) to measure perceived usability in voice systems. This research highlights the importance of combining subjective metrics with objective interaction data to obtain a more complete view of the user experience [9]. However, there are still few studies that integrate advanced deep learning techniques with these scales to generate predictive models [10].

Furthermore, reviews of voice assistants and chatbots have highlighted the relevance of contextual factors such as environment, user familiarity, and task type in the perception of usability [11]. Despite this, most studies continue to address these variables in isolation or through conventional statistical analyses, without leveraging the potential of DNN models to capture complex interactions between them. The gaps identified in the literature include:

- The absence of predictive models specifically designed for Conversational User Interfaces (CUI) that integrate demographic, modality, and subjective perception variables.
- The lack of comparative studies between voice and text modalities using the same structured database, subsequently segmented for independent analysis.
- The limited exploration of the use of automatic usability predictions as a support tool in User-Centered Design.

This work contributes to closing these gaps by proposing an approach based on deep neural networks capable of learning from real user data and predicting the perceived usability of different conversational systems. By doing so, the incorporation of artificial intelligence techniques within the design and evaluation processes is promoted, strengthening the connection between predictive model engineering and the fundamental principles of human centered design.

3 Materials and Methods

For the training and evaluation of the predictive model, the process illustrated in Figure 1 was followed, consisting of five different stages that represent the general development flow for a deep neural network model. This flow begins with data selection and processing, where the most important characteristics are selected and the data is prepared for subsequent analysis. Next, in stage 2, the training and validation set is generated by dividing the data for the purpose of training and evaluating the model. Once the data has been consolidated into training and validation data, the deep neural network is defined and constructed by configuring its architecture. In stage 4, the model is trained using the previously generated architecture and the structured data set. Finally, in stage 5, performance measures are calculated to evaluate the accuracy and effectiveness of the model.

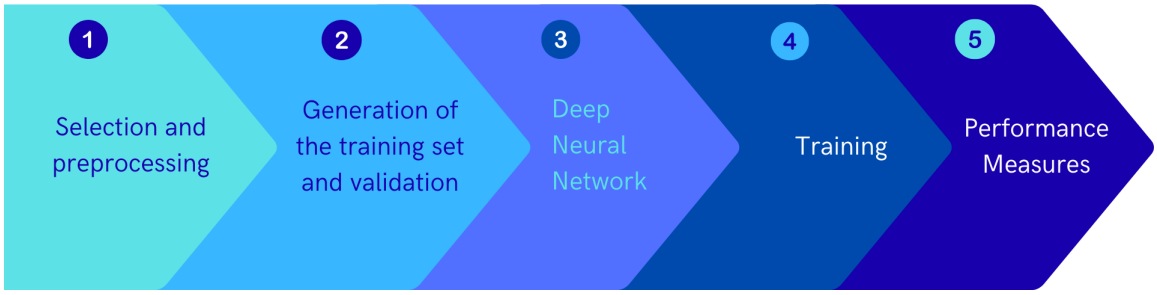

Figure 1: Development workflow of a model based on deep neural networks.

3.1 Selection and Preprocessing

In the data collection process, a sample of 62 participants residing in different departments of Colombia was selected, with ages ranging from 18 to 72 years. The interviews included questions regarding gender, age, ethnicity, and knowledge of the system being evaluated, in both text and voice modalities. All potential user profiles were considered, including individuals with prior experience using CUIs as well as those without such experience. For this purpose, each participant interacted with two systems (Gemini and ChatGPT)in both voice and text modes.

To collect the data, structured interviews were conducted applying the SUS (System Usability Scale) [5] and VUS (Voice Usability Scale) [6] methods for the text and voice modalities, respectively. Each method consists of 10 questions related to the usability of the systems. As mentioned previously, demographic data were also gathered regarding gender, age, municipality, department of residence, ethnicity, and prior knowledge of the system. The age variable is divided into three distinct categories according to [12] . The names of the categories are: "Emerging adulthood," between 18 and 29 years old; "Young and middle-aged adults," between 30 and 39 years old; and "Middle to late adulthood," between 40 and 65 years old. This categorization is carried out to reduce the complexity of the analysis within the neural network. In the "Emerging adulthood" category there are 26 participants, in "Young and middle-aged adults" there are 20 participants, and in "Middle to late adulthood" there are 16 participants.

It is important to mention that the questions are used to calculate the weighted scores for the usability tests in both modalities. This score determines the usability perceived by the user. The aforementioned score is used in the modeling process as an objective variable, that is, the variable that is estimated from the input data. High usability corresponds to a score greater than 80 and low usability to a score less than 80, taking into account what is stated in articles [13] and [14]; the first stipulates a categorization of the SUS scale into Excellent, Good, Marginal, and Poor, where excellent corresponds to a weighted score greater than 80, while the second article states that on the VUS usability scale, the average of user responses to define a system as usable is 80.3. For this reason, in order to facilitate subsequent analysis, it is considered that within the SUS scale and the VUS scale, a score of 80 should be used in order to maintain the same range of measurements on both scales.

It should be noted that, during data collection, a single consolidated database was generated that integrated all the information from the four system and modality combinations. Subsequently, in the preprocessing stage, this database was divided into four independent subsets, corresponding to the Gemini-text,

Gemini-voice, ChatGPT-text, and ChatGPT-voice combinations. Each of these subsets was used to train and validate the respective neural network models, maintaining consistency in the variables and balance in the data distribution. In this way, the coded demographic data and usability scores (SUS and VUS) served as input variables for the deep learning models, in order to predict the level of perceived usability in each system and modality. This facilitates.

3.2 Data Sampling

For the analysis, the database was segmented into four subsets corresponding to the combinations of system and interaction modality: ChatGPT Text, ChatGPT Voice, Gemini Text, and Gemini Voice. Each subset contained the corresponding interaction records along with their respective usability scores derived from SUS for text and VUS for voice. This segmentation allows for comparing and evaluating the performance of the systems under different conditions of use.

The scores obtained were used as a target variable to model the perception of usability, which facilitated the interpretation of results and the application of predictive models. In addition, coded demographic variables were used to identify possible patterns or significant differences in the perception of usability among different user groups. Careful extraction and organization of these samples is important, as it ensures the representativeness and validity of the data for the development and training of subsequent models, allowing for detailed and comparative analysis between systems and modalities.

3.3 Deep Neural Network

A Deep Neural Network (DNN) was used for predictive modeling of perceived usability in CUI systems. This approach allowed complex and nonlinear relationships between coded demographic variables and usability scores (SUS and VUS) to be captured. In addition, the network architecture was designed with multiple hidden layers and activation functions suitable for optimizing the learning of patterns present in the data. Demographic variables and user evaluation of the system/modality were used as inputs, while the output was the perceived usability score.

The model was trained using cross-validation techniques to ensure generalization, and regularization strategies were implemented to prevent overfitting. The network's ability to learn hierarchical representations allowed for more accurate identification of the relevant characteristics that influence the perceived usability of the evaluated systems. This method contributes to the generation of a robust predictive model that can be used to anticipate the perceived usability by different user profiles in conversational systems, facilitating continuous improvement and personalization of the user experience.

3.4 Training and Validation

For model training, a Multilayer Perceptron (MLP) implemented using the Keras Sequential API was used. The architecture consisted of an input layer with a dimension equal to the number of independent variables, followed by two densely connected hidden layers: the first with 64 neurons and the second with 32, both using the ReLU activation function. The output layer consisted of a single neuron with sigmoid activation.

The database was divided into two subsets: 80% for training and 20% for validation. The training process was carried out over multiple epochs, using the Adam optimizer, due to its ability to dynamically adjust the learning rate and its robustness against convergence problems. To avoid overfitting, regularization techniques such as dropout between hidden layers and the early stopping strategy were incorporated, which stopped training if no improvements in validation loss were observed during a certain number of epochs. Throughout training, the loss and mean absolute error (MAE) curves were continuously monitored for both training and validation data. This approach allowed the model to be optimally adjusted, maximizing its generalization capacity and ensuring that the learning was not biased or overfitted to the training sample. The performance metrics obtained subsequently, such as R^2, RMSE, and RME, confirmed the effectiveness of the network in correctly estimating perceived usability levels in different interaction scenarios.

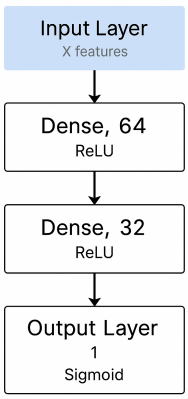

Figure 2: Architecture of the Multilayer Perceptron (MLP) used to predict perceived usability.

3.5 Performance Metrics

To evaluate the performance of the Deep Neural Network model in predicting perceived usability in CUI systems, statistical and graphical metrics widely recognized in the field of machine learning were used.

The quantitative metrics applied included the coefficient of determination (R^2), the mean error (RME), and the root mean square error (RMSE). The R^2 coefficient allowed us to determine the degree of correlation between the values predicted by the model and the actual values, indicating the proportion of variance explained by the model. The MAE provided a measure of the average bias in the predictions, while the RMSE provided a more error sensitive view by penalizing larger deviations between the actual and estimated values. In addition to these numerical metrics, the loss curves and mean absolute error (MAE) curves were analyzed during the model training and validation process. These curves allowed us to observe the learning behavior in each epoch, identifying possible cases of overfitting or underfitting, as well as the convergence of the model towards an optimal solution. Together, these measures provided a comprehensive assessment of the model's performance, allowing us to validate not only its predictive accuracy, but also its stability and generalizing ability across different datasets derived from the systems and interaction modalities.

4 Results

4.1 Text-Based Conversational User Interface

4.1.1 ChatGPT

In the case of the ChatGPT system in text-based mode, the model demonstrated strong performance in predicting perceived usability. A Root Mean Square Error (RMSE) of 0.2 was obtained, indicating that, on average, the model's predictions differ from the actual values by a relatively small magnitude. The Mean Absolute Error (MAE) was 0.07, confirming that the average absolute error between the predictions and the observed values is minimal. Furthermore, the model achieved a coefficient of determination (R^2) of 0.78, suggesting that 78% of the variability in usability levels can be explained by the demographic and experience variables considered. Lastly, the Relative Mean Error (RME) was 0.29, indicating that the mean error represents approximately 29% of the actual value on average. These results reflect a good predictive capability of the model in identifying patterns in the perception of usability for the ChatGPT system in its text-based mode.

Table 1: Model Results with ChatGPT Data for CUI by Text

Metric	**RMSE**	**MAE**	R^2	**RME**
Value	0.20	0.07	0.78	0.29

Fig. 3 shows the training and validation curves corresponding to the evolution of the loss and mean absolute error (MAE) of the developed model for the ChatGPT system in its text mode. In both graphs, a progressive and stable decrease of the metrics can be observed throughout the 50 training epochs, for both the training and validation sets. The loss curve, calculated using mean squared error (MSE), shows a

sustained reduction without abrupt increases during validation, indicating that no overfitting was evident. Similarly, the MAE curve confirms a continuous improvement in the model's performance, with the average error decreasing in both training phases. These visual curves support the quantitative results obtained by the model (RMSE: 0.2, MAE: 0.07, R^2: 0.78, RME: 0.29), demonstrating good generalization ability to predict the system's usability perception.

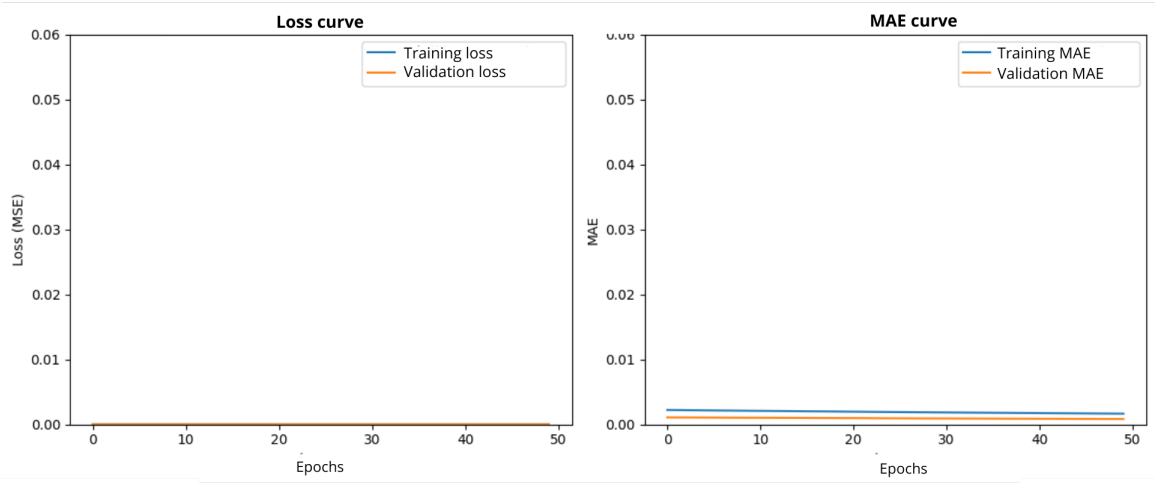

Figure 3: Loss curve and MAE curve of ChatGPT for text modality.

4.1.2 Gemini

In the case of the Gemini system, the model showed a moderate performance in predicting usability perception. The RMSE (Root Mean Square Error) was 0.34, indicating a greater dispersion of errors compared to other evaluated models. The MAE (Mean Absolute Error) reached a value of 0.18, reflecting a higher average error in the predictions. The coefficient of determination R^2 was 0.51, implying that the model explains approximately 51% of the variability in usability perception based on the considered factors. Additionally, the RME (Relative Mean Error) was 0.50, meaning that the average error represents 50% of the actual value, which suggests an acceptable performance but with clear opportunities for improvement.

Table 2: Model Results with Gemini Data for CUI by Text

Metric	**RMSE**	**MAE**	R^2	**RME**
Value	0.34	0.18	0.51	0.50

Fig. 4 shows the loss and mean absolute error (MAE) curves obtained during the training of the model for the Gemini system in text mode. Although both training curves show a decreasing trend, there is a considerable difference compared to the validation curves, which remain high and stable throughout the epochs. This discrepancy suggests that the model did not achieve adequate generalization on the validation data, possibly due to high variability in the responses or insufficient model complexity to capture the underlying patterns. These findings are consistent with the obtained metrics (RMSE: 0.34, MAE: 0.18, R^2: 0.51, RME: 0.50), which demonstrate moderate predictive capacity and the need to adjust or improve the model for this type of data.

4.2 Voice-Based Conversational User Interface

4.2.1 ChatGPT

For the ChatGPT system in its voice mode, the model achieved notable performance in predicting perceived usability. The RMSE was 0.27, reflecting a low standard deviation in prediction errors, while the MAE,

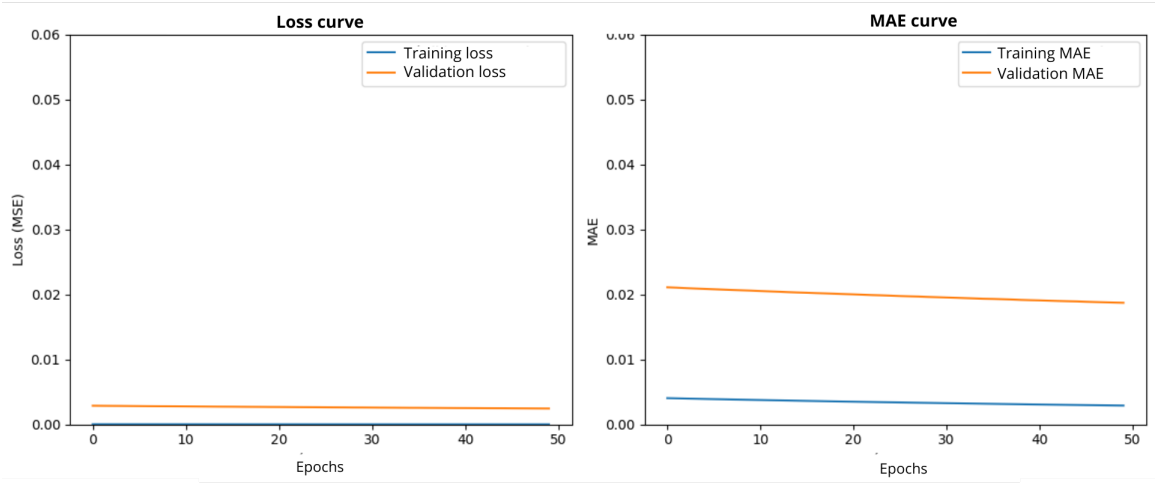

Figure 4: Loss curve and MAE curve of Gemini for text modality.

with a value of 0.13, indicates that the average absolute error remained relatively low. The coefficient of determination R^2 reached 0.71, suggesting that the model was able to explain 71% of the variability in usability perception, demonstrating strong explanatory power. Likewise, the RME was 0.23, indicating that the mean error represents only 23% of the actual value, highlighting the overall precision of the model. Together, these results show that the model offers good predictive capability when working with voice interaction data in the ChatGPT system, making it an effective tool for analyzing user experiences in this modality.

Table 3: Model Results with ChatGPT Data for CUI by Voice

Metric	**RMSE**	**MAE**	R^2	**RME**
Value	0.27	0.13	0.71	0.23

Fig. 5 shows the loss and MAE curves of the model applied to the ChatGPT system in its voice mode. Both graphs reveal a marked difference between the model's behavior on the training set and the validation set. The loss and mean absolute error during training are notably low from the early epochs, suggesting an almost perfect fit to the training data. However, the validation curves present higher and relatively stable values, which may indicate slight overfitting or difficulties for the model to generalize properly to new contexts. Despite this, the obtained metrics (RMSE: 0.27, MAE: 0.13, R^2: 0.71, RME: 0.23) reflect good overall predictive performance, with adequate capability to model the usability perception in this system.

4.2.2 Gemini

In the case of the Gemini system in its voice mode, the model demonstrated high predictive performance in estimating perceived usability. The RMSE was 0.19 and the MAE was 0.06, indicating that prediction errors were low both on average and in dispersion, reflecting high accuracy. The coefficient of determination R^2 reached a value of 0.80, implying that the model was able to explain 80% of the variability in usability perception based on the independent variables, demonstrating excellent explanatory power. Additionally, the RME was only 0.07, indicating that the mean error represents just 7% of the actual value, further reinforcing the model's reliability. Overall, these results highlight the model's robustness for analyzing voice interactions in the Gemini system, even surpassing in some aspects the performance observed in other evaluated configurations.

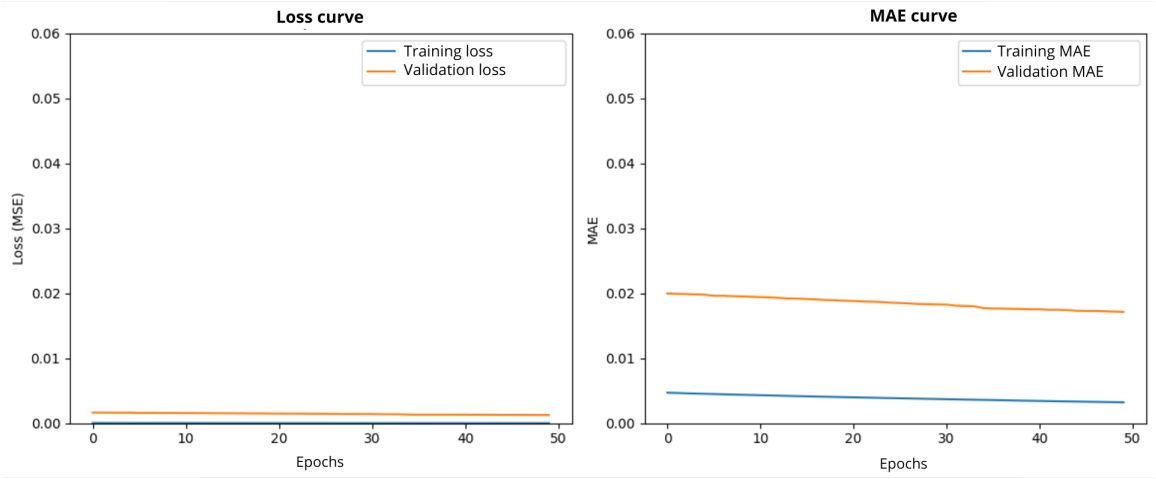

Figure 5: Loss curve and MAE curve of ChatGPT for voice modality.

Table 4: Model Results with Gemini Data for CUI by Voice

Metric	**RMSE**	**MAE**	R^2	**RME**
Value	0.19	0.06	0.8	0.07

The displayed graph shows the behavior of the model during the training and validation process, evaluated using two metrics: loss (using mean squared error or MSE) and MAE (mean absolute error). In the first graph, the training loss is very low and remains almost constant, while the validation loss gradually decreases, indicating that the model is learning properly without signs of overfitting. In the second graph, the MAE also decreases in both training and validation, showing behavior consistent with the loss. These results reflect that the model has good generalization capability, which is confirmed by the obtained metrics: an RMSE of 0.19, an MAE of 0.06, an R^2 of 0.80 (implying that 80% of the variance in the data is explained by the model), and a relative mean error (RME) of 0.07, indicating a low level of relative error. Overall, both the graphs and the metrics suggest a precise and effective model for the regression problem addressed.

5 Discussion

The results obtained through the developed models confirm the effectiveness of the Deep Neural Network (DNN) based approach for predicting perceived usability in conversational systems across different interaction modalities and technological configurations. Based on the general method described in Fig. 2, it was evident that each stage of the process—from data collection and preprocessing to performance evaluation contributed significantly to building robust and generalizable models, especially when considering users' demographic characteristics and prior experience.

The model applied to the ChatGPT system in text mode exhibited solid performance, standing out for its low prediction error (RMSE: 0.2, MAE: 0.07) and a coefficient of determination $\mathbf{R^2}$ of 0.78. This performance indicates that the model was able to explain a considerable proportion of the variance in perceived usability levels, capturing relevant patterns from the encoded variables. The training and validation curves associated with this model showed a progressive decrease in error without signs of overfitting, confirming adequate generalization. In contrast, the Gemini model in text mode showed limited performance ($\mathbf{R^2}$ = 0.51), revealing a considerable gap between the training and validation data. This result suggests potential problems with underfitting or architectural configuration, as the network may not have adequately captured the nonlinear relationships between input variables and perceived usability. Potential causes include insufficient model complexity, a small or undiversified dataset, and excessive regularization that restricts learning capacity. To improve performance, it is proposed to adjust the neural architecture by increasing the depth

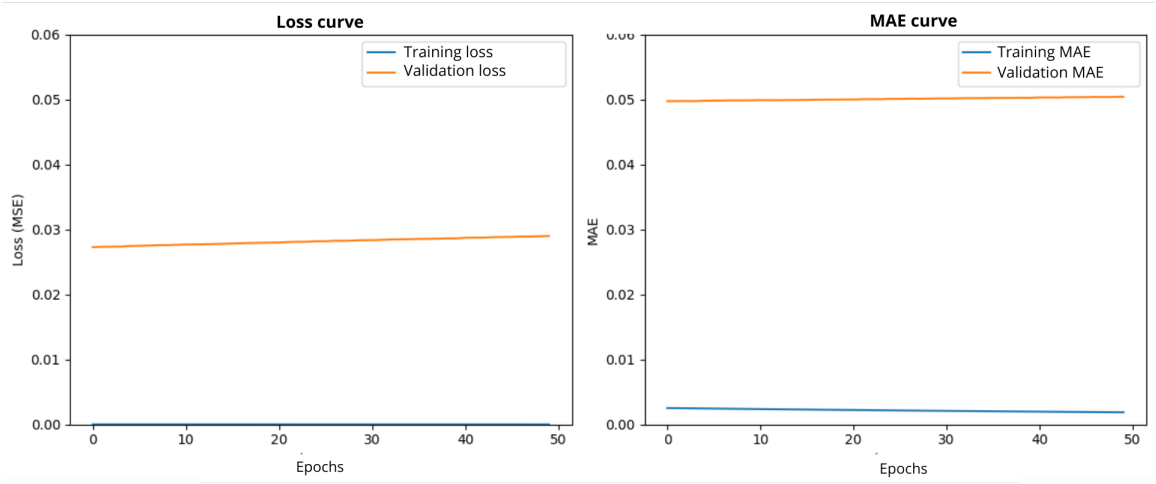

Figure 6: Loss curve and MAE curve of Gemini for voice modality.

or number of neurons, expand the training set by incorporating more representative samples from different user profiles, and optimize the hyperparameters through cross validation. These actions could strengthen the model's predictive capacity and improve its generalizability, thus contributing to a more accurate estimation of perceived usability in conversational interfaces.

Regarding the voice modalities, generally better predictive performance was identified. The model applied to ChatGPT in voice mode obtained satisfactory metrics (RMSE: 0.27, MAE: 0.13, R^2: 0.71), although the validation graphs showed a slight tendency toward overfitting. Nevertheless, the model maintained good predictive capability, demonstrating that voice as an interaction channel can provide more consistent data to model usability perception. The best overall performance was recorded by the model corresponding to Gemini in voice mode, with an RMSE of 0.19, an MAE of 0.06, and an R^2 coefficient of 0.80. These metrics reflect outstanding explanatory and predictive capacity, as well as low dispersion and bias in the predictions (RME: 0.07). The loss and MAE curves during training and validation showed coherent behavior without evidence of overfitting, indicating that the model effectively learned the underlying data patterns. This result suggests that, under adequate training conditions and architecture, DNN models can achieve optimal precision levels even with moderate samples, provided there is proper data representation.

Segmenting the database into four subsets (system and modality) was key to performing a rigorous comparative analysis. This approach revealed that voice modalities tend to offer better predictive results, possibly due to less ambiguity in user perception or a clearer interaction experience. Likewise, the models associated with Gemini in voice modality and ChatGPT in text modality were the most balanced in terms of generalization and accuracy, highlighting that both the type of system and the interaction modality have a significant effect on perceived usability. Taken together, these findings validate the usefulness of the proposed methodological approach, particularly the application of multilayer neural networks with cross validation, regularization, and overfitting prevention techniques. The architectural design adopted with hidden layers, ReLU and sigmoid activation functions proved effective in predicting perceived usability.

Furthermore, beyond quantitative performance, the results of this study offer relevant contributions to the Human-Computer Interaction (HCI) community and to developers of conversational systems. First, the proposed models can be used as predictive tools to support user centered design, enabling early usability assessments without the need for extensive testing. Second, the results provide objective and measurable criteria to guide improvements in conversational interfaces, facilitating data driven decision making. In addition, prediction is useful for early design phases, comparison of alternatives, and cost reduction; this architecture performs well in predicting the usability of voice-based CUI systems. Although this article does not focus on using DNN models to improve systems based on the use of UDC, it does allow us to analyze

how data behaves in a DNN. The main purpose of this article is to show that models based on Deep Neural Networks (DNN) can be used as predictive tools to estimate the perceived usability of conversational systems, based on variables related to the user's profile, their previous experience, and the interaction modality, without the need for extensive traditional evaluations. The results obtained show that, using a properly configured multilayer architecture, it is possible to model nonlinear relationships between these input variables and the subjective perception of usability, achieving accurate and generalizable predictions. In this sense, the proposed approach allows us to anticipate how different system configurations and modalities (text and voice) will be perceived by users, providing objective support for early usability evaluation, comparison between design alternatives, and data driven decision making in the development of conversational interfaces.

6 Conclusions

This research explores the potential of deep learning models, particularly Deep Neural Networks (DNNs), to model the relationship between observable user-related variables (such as demographic factors, previous experience, and interaction modality) and perceived usability in conversational user interfaces (CUIs). The results indicate that DNNs are capable of capturing relevant patterns in usability data obtained through post-test instruments, contributing to a better understanding of how different factors are associated with the subjective perception of usability.

From a methodological perspective, the proposed approach integrates human, contextual, and technological factors into the user experience evaluation process. Rather than replacing traditional User Centered Design (UCD) methods, the use of deep learning based models is proposed as a complementary analytical tool that supports the interpretation of usability data collected using established instruments such as SUS and VUS. The results suggest differences in model performance depending on the interaction modality, with better predictive performance observed in voice-based conversational systems, while clear opportunities for improvement are identified in the textual modality.

As future work, we propose a strategy consisting of collecting usability data, training predictive models, applying these models to new scenarios, and using the resulting estimates to support and prioritize design decisions before full user evaluations. Thus, DNNs act as a complement to User-Centered Design (UCD), enabling early evaluations and cost reduction in the early stages. The results indicate better predictive performance in voice-based conversational systems, while opportunities for improvement are identified in the text modality. As additional future work, we propose expanding and diversifying the samples, comparing DNNs with other machine learning approaches, and incorporating contextual and emotional variables to achieve more comprehensive models.

Acknowledgment

To the LITUX group at the Autonomous University of Zacatecas and the IDIS group at the University of Cauca.

References

[1] L. C. et al., "What makes a good conversation? challenges in designing truly conversational agents," Jan 2019.

[2] C. Murad and C. Munteanu, "'i don't know what you're talking about, halexa': The case for voice user interface guidelines," Aug 2019.

[3] B. R. C. C. Murad, C. Munteanu and L. Clark, "Finding a new voice: Transitioning designers from gui to vui design," Jul 2021.

[4] L. K. G. Samboni, "Revisión de heurísticas y escalas de usabilidad y ux para tecnologías basadas en el habla," 2025.

[5] J. Brooke, "Sus - a quick and dirty usability scale," available: https://www.researchgate.net/publication/228593520$_SUS_Aquick_and_dirty_usability_scale$.

[6] A. M. L. Fulfagar, A. Gupta and A. Shrivastava, “Development and evaluation of usability heuristics for voice user interfaces,” 2021.

[7] E. M. P. Muriel Garrreta Domingo, “Diseño centrado en el usuario.” [Online]. Available: https://openaccess.uoc.edu/server/api/core/bitstreams/7451e4fc-f65e-4606-b58c-0e0161a5f672/content

[8] A. M. D. y R. Chalmeta, “User experience and usability of voice user interfaces: A systematic literature review,” Sep 2024.

[9] C. Arpnikanond, “Voice usability scale: Measuring the user experience with voice assistants,” May 2021.

[10] L. W. y. Z. P. B. Yang, “Measuring and improving user experience through artificial intelligence-aided design,” 2020.

[11] A. S. y Y. Li, “Modeling mobile interface tappability using crowdsourcing and deep learning,” Feb 2019.

[12] E. C. Halloran, “Adult development and associated health risks. journal of patient-centered research and reviews,” Mar 2024.

[13] P. T. K. y. J. T. M. Aaron Bangor, “An empirical evaluation of the system usability scale,” Jul 2008.

[14] M. F. S. W. K. L. E. B. G. M. K.-D. T. Alexandros Bousdekis, Mina Foosherian, “Augmented intelligence with voice assistance and automated machine learning in industry 5.0,” Mar 2025.