https://doi.org/10.58245/ipsi.tir.2502.07

Web Application for Large Language Model-based Diagnostic Analysis of Correlation Maps

Vagač, Michal; and Dudáš, Adam

Abstract: ***The process of diagnostic analysis of data often incorporates human experts for problems outside of computer science and data analysis itself. These experts then interpret and explain the events, trends, and relationships identified in the studied data based on their previous domain knowledge. The main insufficiency of this approach to diagnostic data analysis is the lack of experts or their frequent unavailability, which motivated the previous use of pre-trained large language models as a surrogate for human experts in solving simple domain-specific problems. In this work, the web-focused application for diagnostic analysis of data based on the combination of large language models and correlation analysis is designed, implemented, and examined on two case studies of commonly utilized benchmarking datasets. The proposed model emphasizes the use of interactive visualization techniques in the context of correlation maps, in which the significant relationships between the value of attributes of a dataset are identified, while the large language model offers a brief explanation of these relationships from the point of view of the data domain.***

Index Terms: ***Correlation Maps, Diagnostic Analysis, Large Language Models, Web Services***

Manuscript received February 15, 2025

This research was funded by a grant from the Grant Agency of the Ministry of Education, Science, Research, and Sport of the Slovak Republic (KEGA No. 010TTU-4/2025).

Michal Vagač is with the Department of Computer Science, Matej Bel University in Banská Bystrica, Slovakia (e-mail: michal.vagac@umb.sk). Adam Dudáš (corresponding author) is with the Department of Computer Science, Matej Bel University in Banská Bystrica, Slovakia (e-mail: adam.dudas@umb.sk).

1. Introduction

Traditional diagnostic data analysis heavily relies on human experts to interpret findings and provide meaningful explanations of trends, events, or patterns identified in the data [17, 8]. Since this reliance can be constrained by factors such as the scarcity of qualified experts and the high costs associated with their involvement, the large language models (LLMs) offer a solution to the simplest of these diagnostic problems. With the use of these pre-trained models in processing and understanding vast amounts of information, analysts can automate aspects of data interpretation, which are conventionally slowed down by the lack of human experts.

Correlation analysis is a statistical method relevant to all of the data analysis types. In the diagnostic data analysis, the correlation between the values of a pair of attributes identifies and quantifies the strength and direction of their relationship – also known as prediction potential [19]. These relationships between attribute values are one of the common potential causal factors of phenomena observed in diagnostic analysis. However, without any form of expertise in the area, the analysed data comes from, analysts are hardly able to decide whether the identified relationship is of interest or not [9, 16].

The main objective of the work presented in this study is the design, implementation and experimental evaluation of a web service for basic diagnostic analysis based on the combination of correlation maps and large language models. This approach, as proposed, is composed of the following components:

- The use of correlation analysis for identification of the strength and direction of linear or non-linear monotone functional relationships between attributes of a studied dataset is supplemented by the utilization of a large language model. This large language model serves as a surrogate for human experts and identifies whether the identified relationships are of importance or not with the use of a linguistic description of the relationship.
- With the use of the two aforementioned components – correlation map and large language model – the novel visualization model called Interactive Diagnostic Sheet is constructed. The interactivity mentioned in the title of the proposed visualization model is crucial for both analytical and user experience purposes and includes two interactivity types – hovering interactivity and map element selection interactivity.
- Lastly, the created large language model-based diagnostic analysis of correlation maps and its visualization is presented in the form of a freely available web application.

After the implementation itself, the functionality and diagnostic capabilities of the proposed model are examined through the case studies of two conventionally used benchmarking datasets.

The body of this work is structured into five main sections. Section 2 contains a brief overview of the research works related to the area examined in this article – mainly the application of correlation analysis, diagnostic analysis, and the use of large language models in these analyses. In Section 3, the background, workflow and design of the proposed model are introduced, while the implementation of this design is described in Section 4. The last two sections contain case studies conducted on two benchmarking datasets, present all the functionalities of the proposed model, and conclude the work with the advantages and disadvantages of the approach and several future work areas.

2. Related Works

Since the research presented in this study is focused on the application of correlation analysis methods and large language models in the diagnostic analysis of data, this section of the article presents a brief review of modern relevant works related to these areas of interest.

Even though diagnostic data analysis is one of the basic types of analysis of data, it is often problematic since it relies on human experts from the area in which the analyzed data was collected. Currently, diagnostic analysis approaches are used in several applications such as measuring the impact of digital transformation on agricultural productivity in Azerbaijan [15], where authors of the work conclude, that the utilization of web pages, software-based solutions, and other information and communication technologies in the context of agricultural enterprises helps to increase the productivity of the agricultural field.

On the other hand, authors of [5] process real-time sensor data in the diagnostic analysis process focusing on the oil and wear condition monitoring within the wind turbine gearboxes. When compared to other techniques and methods for the task, the model based on diagnostic analysis outperforms all the conventional approaches.

In [13], the authors focus on the reasoning behind performance changes in a satellite-based rainfall estimation system designed on the bases of Integrated Multi-satellite Retrievals for Global Precipitation Measurement. Via diagnostic analysis, the authors of the work identify a set of systemic errors (detection errors, temporal bias errors, etc.) and demonstrate a strong, systematic dependency of satellite rainfall errors on precipitation event stages.

Similar to diagnostic analysis, correlation analysis and coefficients are often used in various aspects of the analysis of multidimensional data. In [6], the authors propose a novel multimodal weighted canonical correlation analysis model incorporating an attention mechanism for the monitoring of real-time industrial processes. This model assigns weights to data based on probability density estimation, ensuring balanced consideration of past and current

data and, therefore, provides a scalable solution for multimodal environments with sequential data streams.

Authors of [14] present a novel method for the identification of an association between attributes of multidimensional datasets called AssoRep – association-based representation improvement method. The proposed metric is based on the distance correlation method, improved matrix describing the data space, improved attribute representation with the use of aggregation methods, and the mapping of attribute representation over low-dimensional spaces utilizing conventional dimensionality reduction methods.

As mentioned above, this work aims to utilize large language models as a surrogate for human experts in the context of correlation-based diagnostic analysis. This approach to the unavailability of human experts has been implemented in various science areas over the last few years with mixed results. In [3], the authors explore the potential of clinical utilization of two large language models as the replacement of experts from the area of gastroenterology, mainly to assist a clinician with specialized tasks such as patient counselling or management of complex, specialized diagnostic content.

The large language models were also used in the context of extracting crucial information from complexly structured technical documents in the manufacturing industry [10]. In this work the authors design, implement and evaluate a novel specialized large language model combined with task-centric ontologies and knowledge graphs, called OmEGa – Ontology-based Information Extraction Framework for Task-Centric Knowledge Graphs. With the use of this model, authors explore the structural and contextual challenges of the specific documents.

3. Correlation Maps and Large Language Models in the Context of Diagnostic Analysis

The proposed work consists of the utilization of correlation analysis and large language models in the diagnostic analysis of multidimensional datasets and the design, and implementation of a visualization-based web service for this purpose.

In the context of the research presented in this study, correlation analysis focuses on the identification of relationships and trends between values of a pair of attributes, which can be used in the latter predictive analysis tasks [4].

The strength and direction of this functional relationship between a pair of attributes A and B from the common dataset is measured via correlation coefficient, which acquires values from $[-1, 1]$ interval. In the value of a correlation coefficient, the strength of the relationship is represented by the number of the coefficient and the direction of the relationship is represented by the sign of the value – meaning, the further the value of a coefficient is from 0 the stronger the relationship between the selected two attributes is, while positive values denote correlation and negative anticorrelation of the values of the attribute pair [2].

Generally, there are three commonly used correlation coefficients – Pearson, Spearman, and Kendall coefficients – which can be divided into two groups based on the type of relationship measured by them. The first group focuses on the linear relationships between attributes A and B and consists of the Pearson correlation coefficient, which can be computed as [1]:

$$r(A,B) = \frac{\sum_{i=1}^{n}(A_i - \mu(A))(B_i - \mu(B))}{\sqrt{\sum_{i=1}^{n}(A_i - \mu(A))^2}\sqrt{\sum_{i=1}^{n}(B_i - \mu(B))^2}} \quad (1)$$

where A_i and B_i are the i-th measurement of values for attributes A and B and $\mu(A)$ and $\mu(B)$ are the mean values of these attributes.

The second group of correlation coefficients are the coefficients focused on non-linear monotone relationships in data – Spearman and Kendall coefficients. Both of these measures are based on working with rankings of the attribute values instead of the values themselves and therefore can be also called rank correlation coefficients. The Spearman rank correlation coefficient is computed as [2]:

$$\rho(A,B) = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (2)$$

where d_i is computed as $rank(A_i) - rank(B_i)$

and n is the number of measurements in the studied dataset. For datasets of limited size containing non-linear monotone relationships, the Kendall rank correlation coefficient is recommended [4]:

$$\tau(A,B) = \frac{n_c - n_d}{\frac{n(n-1)}{2}} \quad (3)$$

where n_c is the number of concordant pairs of rankings in the dataset, n_d is the number of discordant pairs of rankings in the dataset and n is the number of measurements in the studied dataset.

3.1 Correlation Matrices and Correlation Heatmaps

Besides the large number of measured entities, modern datasets often contain several attributes these entities are composed of. Since the value of the correlation coefficient can be computed between pairs of attributes strictly, the need for summarization of the correlation analysis results of the whole dataset is needed. One such summarization technique is the correlation matrix of the studied dataset, which is cross-indexed by the available numeric attributes. Therefore, for a dataset of n numeric attributes, correlation matrix $\mathbb{C}$ is of size $n \times n$, while each element $\mathbb{C}_{i,j}$ contains the value of correlation coefficient measured between attributes indexing the i-th row and the j-th column (see Rel. 4). Naturally, correlation matrix constructed in this way has the following properties:

- $\forall i \leq n,\ \mathbb{C}_{i,i} = 1$ – the diagonal of the correlation matrix denotes the correlation of the i-th attribute with itself, and therefore is 1 in all cases.
- $\forall i,j \leq n,\ \mathbb{C}_{i,j} = \mathbb{C}_{j,i}$ – since the correlation of the values of attributes $attr_i$ and $attr_j$ hold in both directions, the upper and lower triangles of the correlation matrix are identical.

One can easily imagine a situation in which the input dataset contains a large number of attributes (tens or hundreds), which is common in modern data. Such datasets lead to large correlation matrices, which are hard to read and interpret, and the identification of interesting relationships in such matrices is highly problematic. This motivated visualization of correlation matrices through correlation maps, which represent the value of correlation coefficient through user-defined color palettes (see Figure 1 for a generalized example of such a map).

Even though the correlation map provides higher readability for the correlation analysis results when compared to correlation matrices, in datasets composed of large amounts of attributes the need for the pruning of this matrix arises. In the work [2] authors propose a simple filtering technique for such matrices called σ-masking, which defines the value of σ for correlation matrix $\mathbb{C}$ as:

$$\sigma(\mathbb{C}) = \frac{\mu(|corr(\mathbb{C})|) + max(|corr(\mathbb{C})|)}{2} \quad (5)$$

where $\mu(|corr(\mathbb{C})|)$ is the absolute mean value of correlation matrix $\mathbb{C}$ and $max(|corr(\mathbb{C})|)$ is the absolute maximal value of $\mathbb{C}$. This σ value denotes the lowest significant value of the correlation coefficient, hence:

$$|corr(A,B)| = \begin{cases} sign, & \text{if } |corr(A,B)| \geq \sigma, \\ insign, & \text{otherwise.} \end{cases} \quad (6)$$

where $sign$ denotes significant correlation values and $insign$ denotes insignificant correlation coefficient value between attributes A and B. In this way, only the significant correlation coefficient values ($corr \geq \sigma$) are included in the correlation matrix, which prunes the considered data space and, therefore, offers better interpretability of the correlation analysis results.

3.2 Large Language Models in Diagnostic Analysis of Correlation Maps

The measuring of the values of correlation coefficients in combination with their visualization in a correlation map and incorporation of a large language model form the basis of the diagnostic analysis approach proposed in this work. The workflow designed for the selected problem can

$$
\begin{array}{cccccc}
 & attr_1 & attr_2 & attr_3 & \dots & attr_n \\
attr_1 & corr(attr_1, attr_1) & corr(attr_1, attr_2) & corr(attr_1, attr_3) & \dots & corr(attr_1, attr_n) \\
attr_2 & corr(attr_2, attr_1) & corr(attr_2, attr_2) & corr(attr_2, attr_3) & \dots & corr(attr_2, attr_n) \\
attr_3 & corr(attr_3, attr_1) & corr(attr_3, attr_2) & corr(attr_3, attr_3) & \dots & corr(attr_3, attr_n) \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
attr_n & corr(attr_n, attr_1) & corr(attr_n, attr_2) & corr(attr_n, attr_3) & \dots & corr(attr_n, attr_n)
\end{array}
\tag{4}
$$

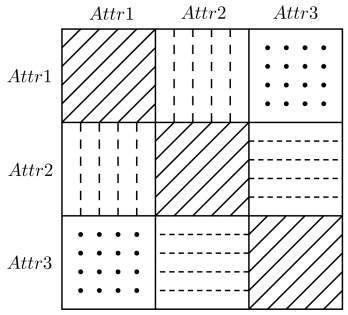

Figure 1: Generalized example of correlation map for a dataset consisting of attributes $Attr1$, $Attr2$, and $Attr3$

be divided into two modules – data topic estimation module and correlation analysis visualization module. Specifically, this workflow has been designed as follows (see Figure 2):

- The header of the input dataset is sent to the large language model for automatic identification of the topic of the dataset, which is crucial from the point of view of the context provided to the large language model in later phases of the process. The first estimate of the topic is presented to the analyst via the user interface:
  - The user can confirm the topic in the case when the large language model estimates the correct area of expertise.
  - The user can correct or complete the original estimation of the model in other cases.

  The correct data topic is then sent back to the large language model to provide context for further interpretation of results.
- The whole dataset is sent to the correlation analysis element of the system, where all the numeric attributes of the dataset are identified and the correlation matrix is built. This correlation matrix is then sent to the σ-masking process, in which the significant values of the correlation coefficient are selected as defined in Rel. 5 and 6. Such a masked correlation matrix is sent to the last – output – component of the proposed system.
- The output of the workflow, which is presented to the user, is comprised of an interactive diagnostic sheet visualization model. This model communicates with the large language model prepared in the previous steps of the method and provides the user with an interactive correlation matrix in which the significant relationships identified in the previous step of the process are highlighted. This interactive diagnostic sheet visualization in its current form offers two types of interactivity:
  - Hovering interactivity – the first of the interactivity functionalities proposed in the visualization model is the hovering above map element interactivity presented in Figure 3. When placing the cursor above a correlation map element, the diagnostic card containing a description of the relationship between the indexing attributes of the element is opened. This description of the relationship is generated by the large language model prepared in the previous steps of the workflow and

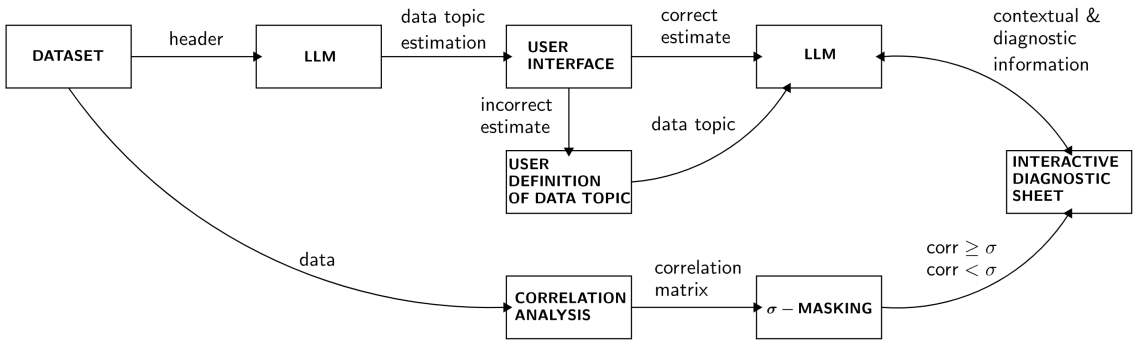

Figure 2: Schema of the proposed LLM-based diagnostic analysis process

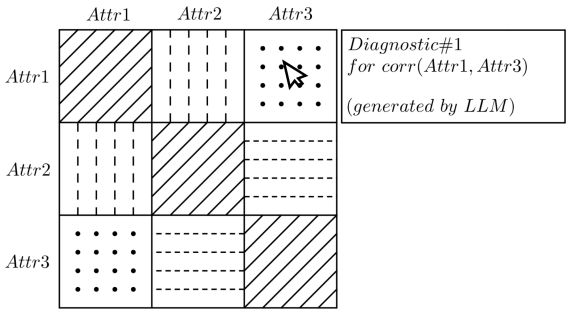

Figure 3: Schema of the proposed hovering interactivity

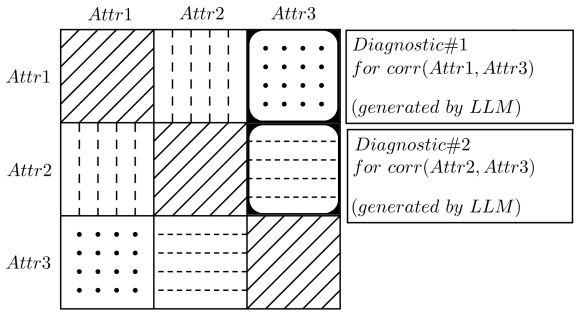

Figure 4: Schema of the selection of a set of map elements interactivity

serves as basic substantiation of the relationship identified in the data.

– Map element selection interactivity – the second interactivity designed for interactive diagnostic sheet visualization is the possibility of selection of a set of correlation map elements, see Figure 4. When clicked, each element of the matrix prompts a large language model and a diagnostic card for the selected pair of attributes is opened. In this way, analysts can examine the causes of relationships between several pairs of attributes of a dataset and identify any similarities or dissimilarities between them.

4. Implementation of Web Service for Diagnostic Analysis

The described system is implemented as a multitier application, consisting of a front-end tier and a back-end tier (see Figure 5).

The front-end tier represents the user interface and is implemented using Angular [11] – a framework for building web applications. As a result, the system is accessible via standard web browsers such as Google Chrome, Firefox or Microsoft Edge.

The back-end tier contains the core logic of the system, in the described system concerning:

- computation of correlation matrix,
- communication with a large language model.

Implementation of the back-end tier follows a standard software architectural pattern, splitting the application between the controller

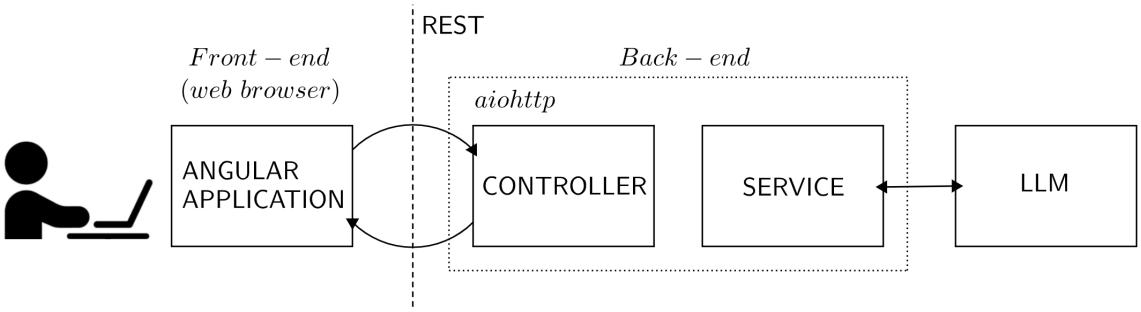

Figure 5: Schema of the utilized architecture

(handling communication with the front-end) and services (handling the work itself). Finally, services are responsible for the implementation of the mentioned core logic of the system.

The sequence of steps conducted by the implemented system is as follows:

1. Dataset upload by a user – the dataset must be a file in a CSV format, which can be zipped, and needs to contain a header for the data. When uploaded, the dataset is stored in the form of a temporary file. In this step, the header of the file is sent to a large language model (OpenAI's $ChatGPT\ 3.5 - turbo$ [18]) for automatic identification of the topic of the dataset. This is implemented using the following prompt formulation:

   ```
   I have the following attributes:
   ${ATTRIBUTES}.
   Print the area of expertise these
   attributes are from.
   ```

   The answer obtained from a large language model is displayed to the user.

2. The user may adjust the identified topic if needed. After that, the analysis can be started. The analysis consists of the computation of the correlation matrix and selecting the significant values of the correlation coefficient, as described above. The attributes pairs with significant values are sent to a large language model for analysis using the following formulation:

   ```
   Is there any known relationship
   between ${ATTRIBUTE_X}
   and ${ATTRIBUTE_Y}?
   ```

   As a result of this step, the user receives a correlation map attributed to the large language model's comments on the significant correlation coefficients.

3. Then, a user can explore a visualised correlation map using the mentioned types of interactivity – hovering interactivity, or the interactivity of selection of an element set from the correlation map. Since this map contains two types of relationships – significant and insignificant – in the implementation of the system, the selection of an element set interactivity has been divided as follows:

   - The significant relationships identified in the previous step of the system can be added to the selected subset via a simple cursor click on the given element of a map.
   - Other than significant relationships are added to the subset via double-click operation.

5. Case Studies on Selected Datasets

After the implementation of the proposed web application, the experimental evaluation of the

designed functionalities is conducted in the context of real-world diagnostic analysis of two conventionally used benchmarking datasets:

- Wine Quality Dataset [12] – dataset composed of 13 attributes measured over 4,898 entities, which measures chemical and qualitative properties (such as fixed acidity of the wine, residual sugars of the wine, density, pH, etc.) of red and white Vinho Verde wine samples from the north of Portugal.
- Appliance Energy Dataset [7] – this data describes appliance energy use in a low-energy building using 19,735 measurements of 29 attributes – temperature and humidity of individual rooms in the building, environmental variables affecting the conditions outside the building, energy consumption, etc.

Since there is no need for further processing of a studied dataset, in the following subsections, two aspects of the proposed system are examined – data topic identification done by a large language model based on dataset header, and interactive diagnostic sheet visualization incorporating the aforementioned interactivity forms.

5.1 Data Topic Idenitification

The first step of the interaction with the proposed system is the upload of the dataset, which needs to be studied. This dataset is, then, sent to the large language model for the task of data topic estimation. Figure 6 presents selected interesting responses of the utilized large language model to this problem. In subfigure (a), the output for a dataset with attributes titled $a, b, c, ...$ is shown – it is natural, that when using such labelling convention, the model is unable to infer the area of expertise from the data correctly. In an ideal case, the query for further specification of the data topic by a user is sent by the system as seen in the subfigure.

In other cases, where the attribute labelling is marginally less vague, the large language model used in the system incorrectly estimates the data topics as shown in subfigures (b) and (c) of Figure 6. These results were both generated from the header of the same dataset – the version of Wine Quality Dataset, where attribute titles were labelled by abbreviations of proper titles of measured chemical and qualitative properties (such as $fA, vA, ...$).

Lastly, subfigure (d) presents the correct estimation of the data topic for the Wine Quality Dataset generated from the header containing full attribute titles – fixed acidity, volatile acidity, and so on.

After the estimation of the data topic, the correlation analysis is conducted and an interactive diagnostic sheet visualization for the input dataset is created. This visualization for the selected evaluation datasets is presented in Figures 7 and 8. In the figures, the relationships of significant strength are highlighted via color saturation settings, eg. in Figure 7 the saturated elements of the map are the following pairs (fixed acidity, citric acid), (fixed acidity, density), (fixed acidity, pH), and (free SO_2, total SO_2).

5.2 Interactive Diagnostic Sheet Visualization

Other than the highlighting of the significance of the relationships identified in correlation analysis shown in Figures 7 and 8, the proposed visualization method offers two types of interactivity presented in Figures 9 and 10.

The hovering interactivity described in the previous sections of this work is featured in Figure 9 on Wine Quality Dataset. As can be seen, the cursor is hovering above the correlation map element describing the relationship between fixed acidity and pH of the wine, while the diagnostic card containing the value of the measured correlation coefficient and description of the relationship generated by the large language model is presented to the user on the right side of the visualization. For lower computational complexity, this interactivity type is active only for the pairs of attributes, whose value of correlation coefficient exceeded the value computed in the σ-masking step of the process.

Figure 10 presents the second type of the proposed interactivity – selection of a set of

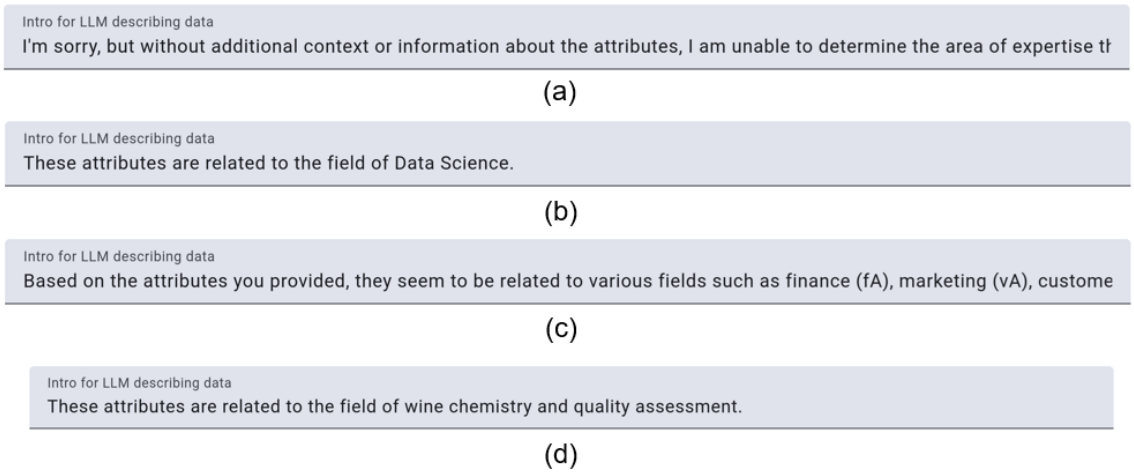

Figure 6: Examples of data topic estimation done in the proposed system

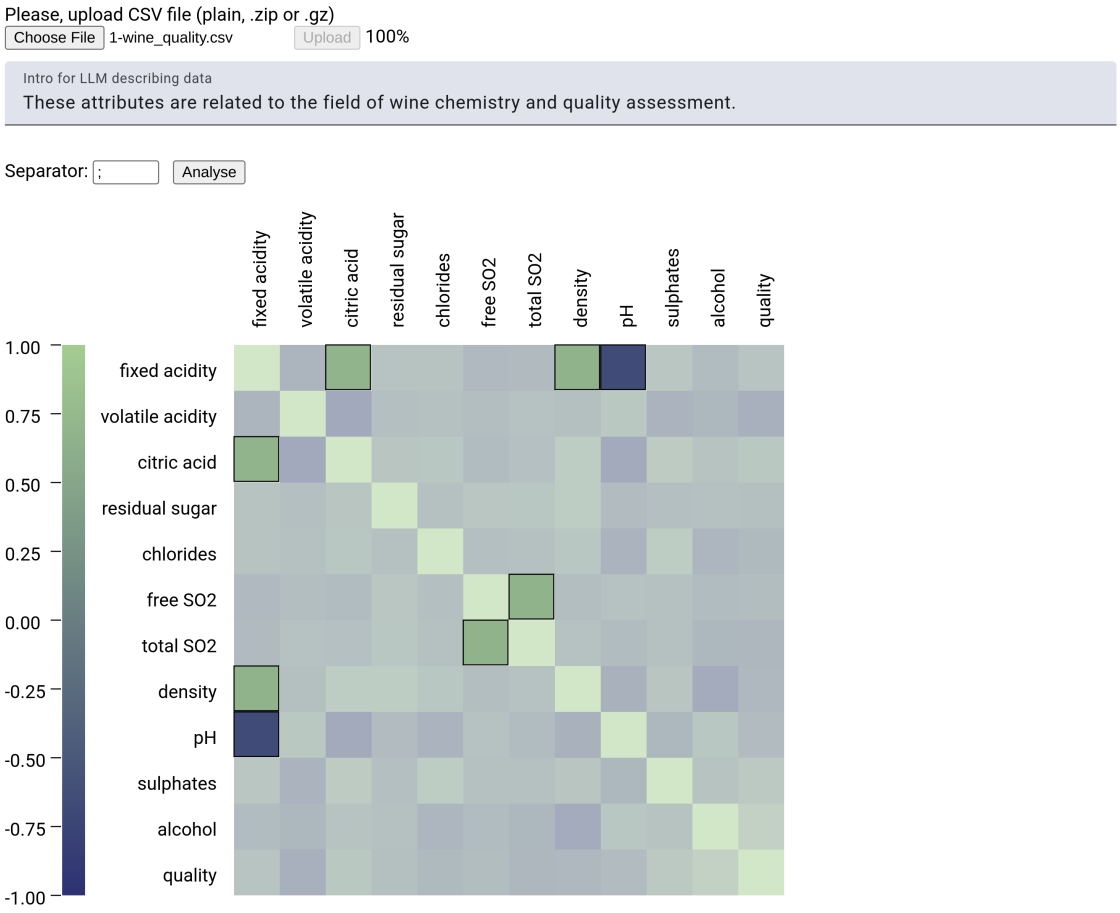

Figure 7: Interactive Diagnostic Sheet Visualization for Wine Quality Dataset

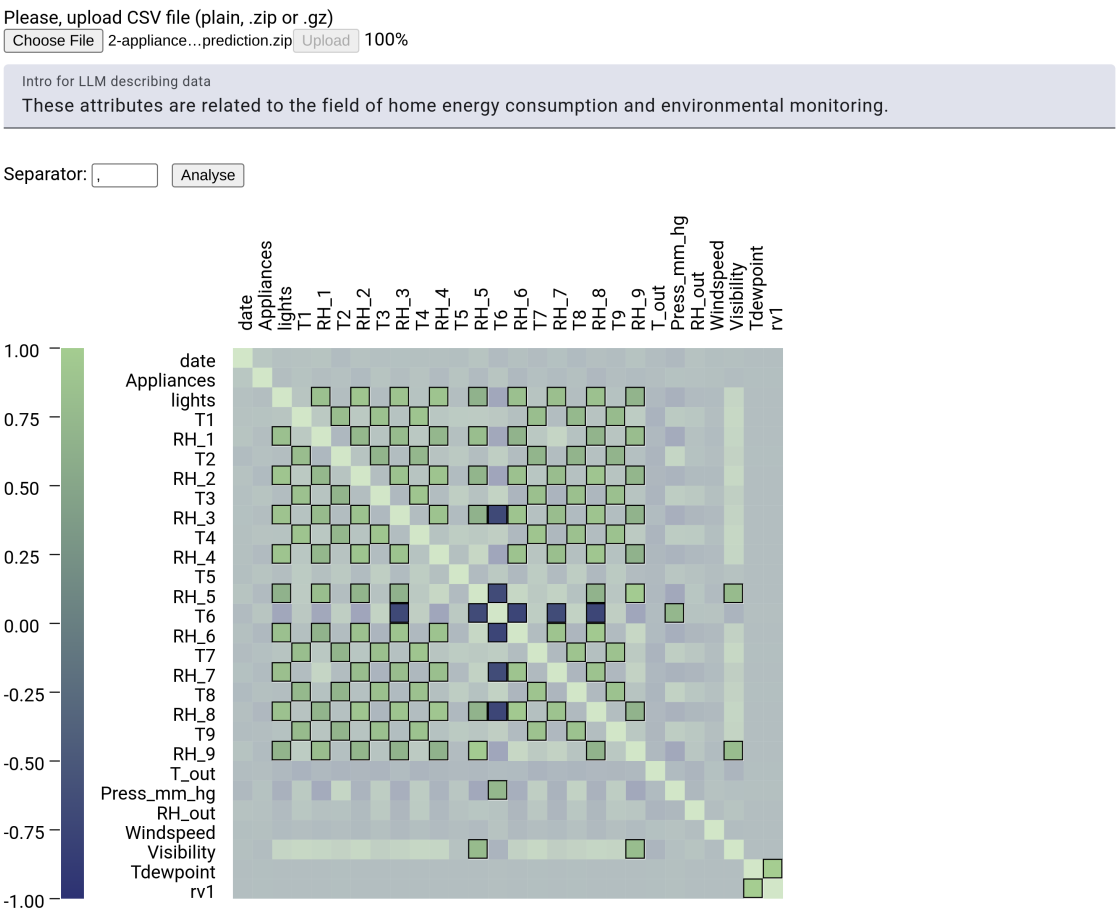

Figure 8: Interactive Diagnostic Sheet Visualization for Appliance Energy Dataset

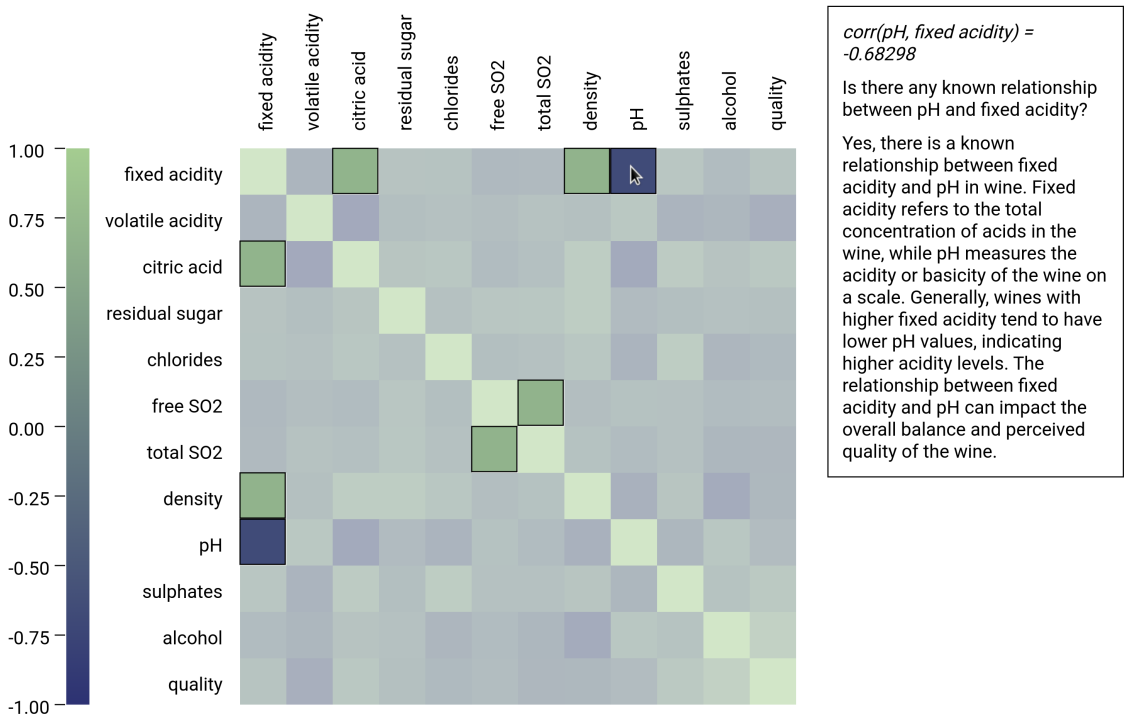

Figure 9: Example of hovering interactivity on Wine Quality Dataset

interesting attribute pairs – on the Appliance Energy Dataset. In the visualization, there are two distinct types of this interactivity based on the strength of the relationship which is being selected:

- The relationships deemed significant in the previous step of the process can be activated by a single click on the map element.
- Other relationships, which did not pass the set σ border, can be added to the selected subset via double-click interaction with the visualization method.

All of the selected relationships between the attribute pairs are highlighted in the visualization by the cross-hatching patterns.

6. Conclusion

This work addresses the disadvantages related to the need for human experts in the process of diagnostic data analysis. Traditional diagnostic methods often require domain expertise to interpret complex relationships within datasets, which can be time-consuming and subjective. To mitigate these challenges, a novel approach that integrates correlation analysis with a large language model was proposed and implemented as a freely available web application. By leveraging the reasoning capabilities of large language models, the proposed system enables automated identification and explanation of significant relationships in the data, reducing reliance on human intuition and expertise.

This diagnostic technique is further enhanced by the development of an interactive visualization model – the Interactive Diagnostic Sheet – which facilitates intuitive exploration and analysis of the identified relationships. This tool allows users to dynamically investigate correlations and their causality, uncover underlying patterns, and refine their understanding of key variables within the dataset, thereby improving the transparency and accessibility of diagnostic data analysis.

Both the application and the proposed diagnostic analysis method were explored us-

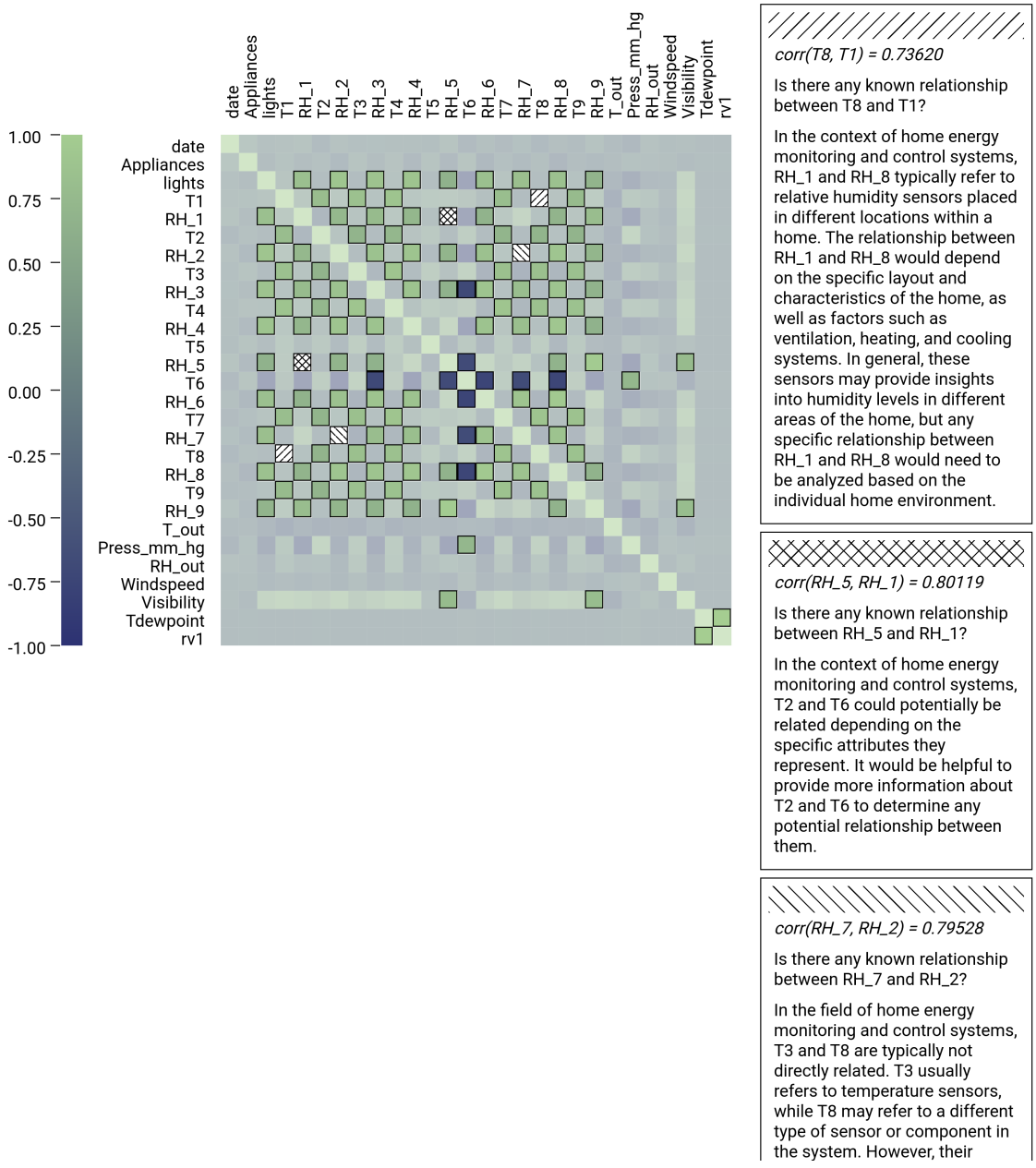

Figure 10: Example of selection of a set of attribute pairs on Appliance Energy Dataset (cropped)

ing two benchmarking datasets – the Wine Quality Dataset and the Appliance Energy Dataset. The results demonstrated that the system successfully identified meaningful relationships, providing insightful preliminary diagnoses. The combination of automated correlation analysis, large language model-based interpretation, and interactive visualization proved effective in assisting users with exploratory data analysis, offering a scalable and adaptable solution for various diagnostic applications.

During the development of the presented model, several future work areas arose:

- Development of advanced interactivity features for the proposed visualization, such as filtering, clustering, and dynamic linking to facilitate more complex and nuanced data exploration.
- Optimization and possible parallelization of large language model prompting for the purposes of highly multidimensional datasets.
- Implementation of ensemble or specialized large language model for higher accuracy of data topic estimation.

Data and Code Availability

The web application presented in this study is freely available at:
`https://chatgpt-correl.vagac.name/`

The datasets used in the experimental evaluation of the application are openly available at:
`https://archive.ics.uci.edu/dataset/186/wine+quality` for Wine Quality Dataset, and `https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction` for Appliance Energy Dataset.

References

[1] A. Dudáš. Correlation n-ptychs of multidimensional datasets. *Lecture Notes in Networks and Systems*, 2024.

[2] A. Dudáš. Graphical representation of data prediction potential: correlation graphs and correlation chains. *Visual Computer*, 2024.

[3] E.J. Gong et al. The potential clinical utility of the customized large language model in gastroenterology: A pilot study. *Bioengineering-Basel*, 2025.

[4] F. Wang et al. Detrended partial cross-correlation analysis-random matrix theory for denoising network construction. *Applied Intelligence*, 2025.

[5] H. Tao et al. A multiple wear sensors online monitoring and warning method in lubricating oil using multidimensional transformer network for wind turbine gearboxes. *IEEE Sensors Journal*, 2025.

[6] J.X. Zhang et al. Multimodal continual learning for process monitoring: A novel weighted canonical correlation analysis with attention mechanism. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[7] L. Candanedo et al. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 2017.

[8] M. Kvet et al. Transaction management in fully temporal system. *Proceedings UKSim-AMSS 16th international conference on computer modelling and simulation*, 2014.

[9] M. Kvet et al. Master index access as a data tuple and block locator. *Proceedings of the 25th conference of Open innovations association FRUCT*, 2019.

[10] M. Shim et al. Omega: Ontology-based information extraction framework for constructing task-centric knowledge graph from manufacturing documents with large language model. *Advanced Engineering Informatics*, 2025.

[11] N. Jain et al. Angularjs: A modern mvc framework in javascript. *Journal of Global Research in Computer Science*, 2014.

[12] P. Cortez et al. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 2009.

[13] R.Z.Li et al. Understanding the error patterns of multi-satellite precipitation products during the lifecycle of precipitation events for diagnostics and algorithm improvement. *Journal of Hydrology*, 2025.

[14] X.Y. Liang et al. A data representation method using distance correlation. *Frontiers of Computer Science*, 2025.

[15] M.K. Gazanfarli. How digitalisation affects agricultural progress in azerbaijan: evidence from panel data approach. *International Joutnal of Sustainable Argicultural Management and Informatics*, 2025.

[16] M. Kvet. Relational data index consolidation. *Proceedings of the 28th conference of Open innovations association FRUCT*, 2021.

[17] S. Masís. Interpretable machine learning with python. 2023.

[18] W.T. Sewunetie and L. Kovacs. Exploring sentence parsing: Openai api- based and hybrid parser-based approaches. *IEEE Access*, 2024.

[19] S.S. Skiena. The data science design manual. 2017.

Michal Vagač is an assistant professor at the Department of Computer Science, Faculty of Natural Sciences of Matej Bel University in Banská Bystrica. His research is focused on computer vision, computer graphics, and robotics and was published in over 45 scientific works.

Adam Dudáš is an assistant professor at the Department of Computer Science, Faculty of Natural Sciences of Matej Bel University in Banská Bystrica. He is the author and co-author of more than 35 research works and his research activities are related to descriptive, explorative and predictive data analysis with emphasis on visualization in the context of statistical analysis of data.