Applied Science and Biotechnology Journal for Advanced Research e-ISSN: 2583-553X
Peer Reviewed Journal
Volume-3, Issue-5 (September 2024), Page: 8-19 https://doi.org/10.5281/zenodo.13805421

Evaluating the Role of Large Language Models Detection: A Comparative Analysis of Noninvasive Testing Methods and AI-Generated Diagnoses

Yue Zhu[1*], Ziwei Wang[2], Xiaoyi Zhang[3], Yuchen Zhang[4] and Jiaqi Hong[5]
[1]Georgia Institute of Technology, USA
[2]The Chinese University of Hong Kong(Shenzhen), China
[3]Jacobi Medical Center/ Albert Einstein College of Medicine, USA
[4]Mailman School of Public Health, Columbia, USA
[5]China Academy of Art, China

[*]**Corresponding Author:** Yue Zhu

Received: 05-08-2024 Revised: 24-08-2024 Accepted: 15-09-2024

ABSTRACT
Nonalcoholic fatty liver disease (NAFLD) has become a global epidemic. The coexistence of NAFLD and type 2 diabetes mellitus (T2DM) is common, and their interaction significantly heightens the risk of adverse clinical outcomes. Despite advancements in medicine, diagnosing NAFLD remains a critical challenge. Large language models (LLMs) have shown exceptional capabilities in various medical applications. However, their potential in diagnosing NAFLD has yet to be fully explored.

Keywords: llm, testing method, ai

I. INTRODUCTION

Nonalcoholic fatty liver disease (NAFLD) is a major global health challenge with its prevalence ranging between 20% to 35% worldwide. Approximately one-third of NAFLD patients may develop nonalcoholic steatohepatitis (NASH), a condition associated with severe complications such as cirrhosis, hepatocellular carcinoma, and higher liver-related mortality [1]. Even without NASH or advanced fibrosis, individuals with NAFLD often linked with rising rates of obesity and metabolic syndrome, face an elevated risk of overall mortality compared to the general population [2-4]. However, a substantial portion of NAFLD patients remain undiagnosed. Liver biopsy is the gold standard for diagnosis of NAFLD and NASH, but the procedure is expensive and has increased risk of complications. Given the multifactorial and intricate etiology of the disease, it's still difficult to determine a specific prevention strategy and achieve early identification of high-risk groups to reduce the prevalence of the disease. Therefore, improved prediction of the risk of NAFLD may be of great value in the prevention and control of the disease in the general population.

In recent years, advancements in artificial intelligence (AI) have opened new frontiers in healthcare, offering innovative solutions to longstanding challenges [5, 6]. Large Language Models (LLMs), such as ChatGPT-3.5 and ChatGPT-4, represent a significant leap in this direction. (https://openai.com/blog/chatgpt). These models have shown exceptional capabilities in generating human-like text, and their potential application in medical diagnostics and risk prediction is a subject of considerable interest. Although LLMs have demonstrated impressive capabilities, attempts to assess the clinical knowledge of models typically rely on automated evaluations based on limited benchmarks. Limited quantitative evaluation of their performance and accuracy has been conducted in specific medical tasks [7].

The potential of Large Language Models (LLMs) in managing Nonalcoholic Fatty Liver Disease (NAFLD) goes well beyond diagnosis. These models are adept at analyzing complex data to uncover subtle patterns and risk factors that traditional methods may miss. This is crucial for NAFLD, where disease progression varies widely among individuals. By using LLMs, we could more accurately predict how the disease will progress, allowing for more personalized and effective treatments, ultimately improving patient outcomes. Additionally, LLMs could significantly enhance patient education and engagement by providing clear and relatable advice, encouraging better lifestyle choices that are key to managing NAFLD. Furthermore, the flexibility of Large Language Models (LLMs) extends to their ability to process various types of inputs, including both textual case descriptions and medical images. This versatility makes LLMs even more accessible and user-friendly for healthcare professionals and patients.

However, the application of LLMs in NAFLD isn't without its challenges. The inherent variability in their responses, a result of their probabilistic design, poses questions about their reliability and accuracy in a clinical setting.

This is a significant concern for NAFLD, where understanding the nuances of a patient's history and lifestyle is essential for proper diagnosis and treatment. Thus, there's a critical need for thorough testing and validation of LLMs in both clinical trials and real-world scenarios to confirm their effectiveness and safety in treating NAFLD. Moreover, as LLMs become a more common tool in healthcare, we must ensure they are used ethically and transparently, avoiding biases, especially since NAFLD affects diverse(sub)groups e.g, DM or non-DM, differently.

As of now, no research has evaluated ChatGPT's accuracy and comprehensiveness when addressing questions specific to NAFLD. Thus, our study aims to explore the feasibility of using LLMs for early detection of NAFLD and compared their performance with that of conventional models, such as the Fatty Liver Index (FLI) and the (USFLI). We further explored GPT-4V's potential in clinical diagnosis usingUltrasound images.

II. METHODS

Study Population

The National Health and Nutrition Examination Survey (NHANES) is a comprehensive cross-sectional research program conducted by the Centers for Disease Control (CDC) and Prevention of the USA whose purpose is to evaluate the health status in the U.S. population after survey weighting by using interview, examination, dietary and laboratory data. The original survey protocol was approved by the Institutional Review Board of the National Center of Health Statistics. All participants signed informed consent forms. The current study was deemed exempt bythe Institutional Review Board of our center.

Data for this research were sourced from the NHANES 2017-2018, with a total of 9,254 participants. The current research employed the controlled attenuation parameter (CAP) for NAFLD diagnosis. Exclusion criteria included: 1) age <20 years, 2) without complete components on FLI, USFLI and FIB-4 calculation, 3) lack of baseline median stiffness, median CAP or with incomplete elastography exam status, pregnancy, 5) excessive alcohol consumption (alcohol consumption >20g/day for males and 10g/day for females), 6) other existed liver conditions, including viral hepatitis infection (defined as a positive HCV RNA, HCV-antibody or HBsAg test), autoimmune hepatitis and liver cancer, and 7) missing follow-up information. Ultimately, 1,542 participants were included in the final analysis. (Figure 1) Definition of NAFLD, Fibrosis, Comorbidities and Index Calculation.

NHANES 2017-2018 employed vibration controlled transient elastography (VCTE) to assess liver fibrosis by measuring liver stiffness and quantifying liver fat through the controlled attenuation parameter (CAP). The device calculated the median liver stiffness measurement (LSM) and CAP along with the interquartile range (IQR). Exams were considered complete if performed in fasted state (>3h) with ≥10 LSM, and a stiffness IQR/median <30% [8]

Fatty liver disease (FLD) was defined as having a median CAP score of 285 dB/m for detecting hepatic fibrosis [9]. The definition of NAFLD was FLD without other causes of chronic liver disease or excessive alcohol consumption (alcohol consumption <20 g/day for males and 10g/d for females). Significant fibrosis and advanced fibrosis were defined as LSM ≥ 8.0 kPa and ≥ 13.1 kPa, respectively.

Moreover, the study also employed FLI and USFLI for identifying individuals with FLD, while FIB-4 was utilized to stratify the risk of advanced fibrosis among FLD. FLI and USFLI are identified as the foremost validated diagnostic panels for the identification and comprehensive assessment of NAFLD among the array of validated diagnostic tools available. The FLI score was computed using parameters including triglycerides, BMI, GGT and waist circumference, whereas the USFLI formula encompassed age, ethnicity, GGT, waist circumference, fasting glucose and fasting insulin. The FIB-4 index, well-validated and widely employed in clinical settings, was utilized to identify participants with advanced fibrosis [10, 11].

FLI and USFLI were calculated using the following formula as previously described:

$$\mathrm{FLI} = \frac{e^{(0.953\cdot\ln(\mathrm{TG})+0.139\cdot\mathrm{BMI}+0.718\cdot\ln(\mathrm{GGT})+0.053\cdot\text{waist circumference}-15.745)}}{1+e^{(0.953\cdot\ln(\mathrm{TG})+0.139\cdot\mathrm{BMI}+0.718\cdot\ln(\mathrm{GGT})+0.053\cdot\text{waist circumference}-15.745)}} \times 100$$

Individuals were defined as FLD if their FLI score ≥ 60 or USFLI score ≥ 30. Advanced fibrosis was assessed by serological non-invasive fibrosis index (FIB-4), calculated using the following formula:

FIB-4 = (age*AST) / (Platelet counts * (SQRT(ALT)))

Participants with FIB-4 > 2.67 were regarded as high-risk for advanced fibrosis. Individuals meeting one or more of the following criteria were considered to have diabetes: 1) fasting plasma glucose ≥ 7.0mmol/L or 2-h oral glucose tolerance test level of ≥ 11.1mmol/L; 3) HbA1c ≥6.5%; 4) use of diabetes medication or insulin; and 4) self-reported doctor diagnosis of diabetes. Cardiovascular disease was defined by self-reported medical history of congestive heart failure, coronary artery disease, heart attack or stroke.

Age, sex, race, BMI, waist circumference, hip circumference, weight, pulse, blood pressure, CAP, LSM, smoking status, drinking status and medical conditions were adopted from demographic and health questionnaires and physical examination of the NHANES survey. Race was classified as non-Hispanic Black, non-Hispanic White, Mexican American, other Hispanic and other races. BMI was calculated as weight in kilograms divided by height (m) squared. Smoking status was categorized as never a smoker (defined as smoking less than 100 cigarettes in life), former smoker (smoking more than 100 cigarettes but smoking not at all now) and current smoker (smoking more than 100 cigarettes and smoking some days or every day now). The drinking status was categorized as either abstinent or consuming alcohol. Waist-to-hip ratio was calculated as waist circumstance divided by hip circumstance. Fasting plasma glucose, HbA1c, 2-h oral glucose tolerance test, serum insulin, total cholesterol, triglycerides, high density lipoprotein, low density lipoprotein, total protein, red blood cell counts, white blood cell counts, monocytes counts, platelet counts, hemoglobin, high-sensitive C-reactive protein, total bilirubin, albumin, aspartate aminotransferase (AST), alanine aminotransferase (ALT), gamma-glutamyltransferase (GGT), alkaline phosphatase, serum uric acid, serum creatinine and urine albumin-creatinine ratio were obtained from laboratory test results. The triglyceride-glucose index is calculated by the formula: Ln [TG (mg/dl) × glucose (mg/dl)/2].

The Prediction of NAFLD Estimate by LLMs (GPT-3·5/GPT-4)

To predict the risk of NAFLD and its associated, we utilized Large Language Models (LLMs) i.e., ChatGPT-3.5 and GPT-4. Before feeding data into these models, we transformed variables into sentences, as shown in Figure X. This transformation aligns with the language-based nature of LLMs, allowing us to receive outputs as individual risk percentages instead of broad text descriptions. Moreover, we included each participant's details, such as age, gender, diabetes status, hypertension, smoking habits, and other variables. Regular expressions were used to extract probabilities related to NAFLD. We also explored the clinical relevance of GPT's probability estimations. To process large volumes of data efficiently, we used the ChatGPT API (GPT-3.5-turbo and GPT-4 as of November 2023) in a Python environment. Furthermore, we present two cases with referenced NAFLD diagnosis with ultrasound images, sourced from The First Affiliated Hospital of Hebei North University employing GPT-4 Version (GPT-4V) to examine the proficiency in medical diagnosis. Both patients signed informed consent forms. GPT-4V, since its introduction in September, 2023, has garnered significant interest due to its capability to process multimodal inputs. For each case, we provided GPT-4V with a prompt to generate a comprehensive clinical report based on ultrasound images of the liver, gallbladder, spleen, pancreas, and portal vein, including details of any lesions observed. The generated reports by GPT-4V were then meticulously analyzed to identify clinically relevant information. Key findings and notable excerpts from these AI-generated clinical reports are highlighted and discussed in Figure 3. This exploration serves not only to assess the current capabilities of GPT-4V in interpreting complex medical imaging but also aims to shed light on the potential of AI-assisted diagnostics in the field of radiology and hepatology.

Comprehensive Evaluation of LLM

A thorough evaluation of LLM for gastroenterological decision support was conducted in the current study, utilizing a dataset of 100 patient cases. Four experienced gastroenterologists with 5 to 15 years of clinical experience assessed the performance of LLM across five dimensions: accuracy, clinical relevance, overall utility, explainability and consistency. They systematically reviewed patients' profiles and reference results, assigning scores to GPT-3.5 and GPT-4.0 on a scale of 1 to 5. A score of 5 denoted alignment with clinical references, 3 indicated the presence of irrelevant content, and 1 signified incorrect suggestion. Additionally, our evaluation was expanded to encompass direct feedback from 20 patients, including ratings based on symptom description accuracy, cost of care, and time efficiency on a scale of 1 to 3 (poor, acceptable and great) associated with the treatment process.

III. STATISTICAL ANALYSIS

All statistical analyses were performed in R (version4.2.1) software ("survey" packages in R account for the complex survey design were used) and Python's SciPy library (version 1.12.0). Sampling weights were considered during analysis. Continuous variables and categorical variables were described as weighted means (95% confidence interval, CI) and frequency counts (percentages), respectively. The comparison among groups were evaluated utilizing one-way ANOVA, and the Pearson chi-square test for categorical variables. Receiver-operator characteristic curve (ROC) analysis was conducted to compare the accuracy of FLI, USFLI, GPT-3.5 and GPT-4.0 for NAFLD prediction. Furthermore, the relationship between different scoring systems were assessed by plotting scatterplots and calculating Pearson's correlation coefficient.

IV. RESULTS

In total, 1,542 participants were enrolled in the present study (detailed inclusion was shown in Figure 1). Over a median follow-up of 23.78(13.00, 35.00) months, 23 (1.30%) patients experienced all-cause mortality, with 2 (0.37%) deaths attributed tocardiovascular causes.

Characteristics of the Study Population

Of the study cohort, 586 individuals were identified as NAFLD. Notably, participants with NAFLD tended to be male and older. In comparison to patients without NAFLD, those with NAFLD exhibited a higher prevalence of cardiovascular disease and diabetes and had a higher possibility of obesity, including elevated BMI, waist-to-hip ratio. Additionally, NAFLD patients demonstrated elevated levels of hepatic inflammatory markers, including AST, ALT, GGT, WBC, monocytes and hs-CRP. Other laboratory indices such as FPG, HbA1c, serum insulin, TyG, TRIG, RBC, hemoglobin, PLT, ALP, serum uric acid and UACR showed significant elevations in NAFLD patients, while HDL and ALB were relatively lower.

Comparison of Risk Scores and LLM on NAFLD Diagnosis

The performance comparison of FLI, USFLI, GPT-3.5 and GPT-4.0 in NAFLD diagnosis were demonstrated in Table 2. The study employed CAP measured by VCTE for final NAFLD diagnosis, while we also utilized scoring systems mentioned above to assess whether participants in NHANES were NAFLD or not. As depicted in Table 2, the FLI score had the highest F1-score of 0.66 (95% CI 0.63-0.69), with a sensitivity of 0.66 (95%CI 0.62-0.70) and specificity of 0.79 (95%CI 0.77-0.82), while GPT-3.5 showed the lowest F1-score, and its sensitivity and specificity were
0.56 (95%CI 0.44-0.58), 0.75 (95%CI 0.78-0.87), respectively. Furthermore, the four methods exhibited varying AUC values, ranging from 70.7% to 83.1%. Notably, specific AUC values of GPT-3.5, GPT-4.0, FLI and USFLI were 70.7%, 83.1%3, 81.7%, 82.7%, respectively (Figure 3). In combination of AUC, F1-score, sensitivity and specificity, the GPT-4.0score demonstrated superior performance in NAFLD diagnosis.

In our subgroup analysis focusing on the impact of Diabetes Mellitus (DM) on NAFLD diagnosis, we observed significant differences in performance between models in populations with and without DM. GPT-4.0 displayed notable adaptability, with an AUROC of 0.825 (95% CI 0.784-0.866) in non-DM individuals and a decrease to 0.749 (95% CI 0.621-0.878) in those with DM, yet it showed a remarkable increase in sensitivity from 0.502 to 0.840. Conversely, GPT-3.5's performance dropped more sharply in the DM group, from an AUROC of 0.707 (95% CI 0.655- 0.758) in non-DM to 0.538 (95% CI 0.359-0.717), indicating a more significant impact of DM on its diagnostic capability. The FLI and USFLI indices also showed variability with DM presence, but the contrast in performance between GPT-3.5 and GPT-4.0 across DM and non-DM groups underscores the importance of considering DM as a factor in NAFLD diagnostic models. GPT-4.0's ability to maintain higher specificity and achieve a substantial sensitivity boost in DM patients highlights its potential for clinical application in diverse patient populations. This analysis underscores the critical need for model adaptation to patient subgroups, especially inconditions like DM that markedly influence NAFLD diagnosis.

Correlation Analysis for Different Scoring Systems

To evaluate the correlations among varied scoring systems, Pearson correlation analysis was performed (Figure X). In the current analysis, the Pearson correlation coefficient demonstrated a significant association between the GPT-4.0 score and the FLI, with the highest Pearson's r value of 0.71 ($P < 0.001$). Furthermore, the correlation between GPT-3.5 and GPT-4.0 was determined to be 0.61, representing the lowest among all analyzed relationships. Moreover, the Pearson's r value between GPT-4.0 and USFLI, as well as FLI and USFLI was 0.71 and 0.64, respectively, indicating a robust and moderate correlation.

Physician-led Comprehensive Evaluation of LLM (GPT-4.0 and GPT-3.5) in NAFLD Diagnosis

Four gastroenterologists assessed the performance of GPT-4.0 and GPT-3.5 for NAFLD diagnosis in diverse dimensions through the evaluation of 100cases randomly selected in this study cohort. The assessment revealed that GPT-4.0 achieved higher scores in accuracy (4.75 vs. 4.0), clinical relevance (4.50 vs. 3.92), overall utility (4.25 vs. 3.75), explainability (4.25 vs. 3.25), and consistency (4.35 vs. 3.65) in comparison to the performance of GPT-3.5. Additionally, this study compared the performance of GPT-3.5, GPT-4.0 and real-world clinical scenarios based on patient feedback, including symptom description accuracy, cost of care, and time efficiency (Figure XC). The scores of symptom description accuracy exhibited a notable increase from GPT-3.5 (1.8) to GPT-4.0 (2.6), approaching to the clinical reference (2.75). Regarding cost of care ratings, high scores were consistently observed among all evaluated entities, where GPT-3.5 received a rating of 2.7, and both GPT-4.0 and the clinical reference achieved a score of 2.8. However, discernible variability in time efficiency was evident, with GPT-4.0 (2.9) displaying enhanced performance compared to GPT-3.5 (1.6).

Comparison between GPT-4V and Radiologist for Liver Ultrasound Analysis

To evaluate the imaging diagnostic efficacy of GPT-4V in medical applications, this study compared the analysis of liver ultrasound conducted by GPT-4V and a radiologist for two cases diagnosed with NAFLD (Figure X). The

findings of the comparative analysis demonstrated that the GPT-4V proficiently identified normal liver parenchyma echotexture and recognized the unremarkable features in the portal vein and common bile duct. Additionally, it effectively detected abnormalities, including an enlarged gallbladder, wall irregularities and the presence of strong echoes with posterior acoustic shadowing. However, GPT-4V failed to accurately quantify precise parameters, such as wall thickness and the enhancement of hepatic parenchymal echogenicity. Therefore, for hepatic ultrasound imaging of NAFLD patients previously diagnosed by an experienced radiologist, GPT-4V exhibited limitations in the identification and diagnosis of NAFLD.

V. DISCUSSION

This study compared the performance of LLMs in NAFLD and fibrosis progression prediction with that of FLI as well as USFLI and validated the output using real-word data. The findings of this study are as follows. GPT-4 achieved performance comparable to the FLI in NAFLD prediction in the NHANES cohort {Specificity (0.853 vs 0.794) and F1 score (0.637 vs 0.660)}and USFLI {Specificity (0.853 vs 0.917) and F1 score (0.637 vs 0.580)}.

Since its release, ChatGPT has attracted considerable attention worldwide because of its exceptional ability to generate plausible responses across various topics. In some cases, ChatGPT has outperformed existing prediction models, encouraging studies on the potential of ChatGPT for use in various applications [14]. For instance, ChatGPT has showed potential in various medical applications such as answering United States Medical Licensing Exam (USMLE) questions [15], answering ophthalmology queries [16], and generating simplified radiology reports for patients [17]. However, limited studies have been conducted on the use of language models for early detection of NAFLD. To the best of our knowledge, this study is the first to reveal that ChatGPT exhibited performance comparable with the conventional risk score model in predicting non-alcoholic fatty liver disease using large real-world medical data. These findings provide insights into the potential applicability of ChatGPT in medical practice.

Our methodological approach compared the performance of LLMs with both FLI and USFLI for NAFLD diagnosis and the FIB-4 index for advanced fibrosis assessment, contributed to the accuracy and reliability of our diagnostic framework of LLMs. We revealed that GPT-4.0 displayed moderate accuracy in the primary diagnosis, indicating its potential usefulness as an adjunct for clinical diagnosis. ChatGPT could generate potential diagnoses based on patient symptoms, medical history, and laboratory test results [18]. Additionally, the results suggested a potential refinement in GPT-4.0's comprehension and processing of medical information compared to GPT-3.5. GPT-4.0 manifests profound grasp of input context, consequently enhancing the accuracy of the generated text. In terms of the output process, GPT-4.0 provided disease diagnoses accompanied by comprehensive explanations. And GPT-4.0 generally answered quicker than GPT-3.5, possibly due to the hardware enhancements and algorithmic improvements [19].

Type 2 diabetes and impaired glucose tolerance are important risk factors for the development of NAFLD and NASH [20, 21]. A larger 2018 meta-analysis of nearly 300,000 individuals from 19 studies indicated that individuals with NAFLD have a greater than 2-fold increased risk for developing incident T2DM [22]. Considering the higher prevalence of NAFLD in patients with T2DM, we also investigated the difference of NAFLD diagnosis among individuals with T2DM and non-T2DM in the same cohort. Our research illustrated attenuated specificity among all models but higher sensitivity in T2DM individuals compared to non-T2DM. This might suggest the potential shared risk factors among NAFLD and T2DM. To be noted, GPT-4.0 demonstrated the highest sensitivity (0.840) and F1 score (0.808) in T2DM group compared to general population and non-T2DM group. Despite the high coexistence of NAFLD and T2DM and increased risk of cardiovascular sequela among these individuals, evidence does not currently support screening for NASH or NAFLD in patients with T2DM. In its 2017 practice guideline, the American Association for the Study of Liver Diseases did not recommend routine NAFLD screening for high-risk patients in primary care, diabetes, or obesity clinics [23]. Given the current understanding of the pathophysiological interplay between insulin resistance in both T2DM and NAFLD.

We also investigated the feasibility of using a large language model LLM, specifically GPT-4V, to extract structured data from unstructured pathology reports in a zero-shot approach. The results of our experiments suggest that GPT-4 can be used effectively to extract relevant information from histopathological reports with high accuracy. This capability holds potential to reduce the workload of human experts while preparing ground truth data for machine learning applications.

This study has also evaluated the performance of LLMs such as GPT by asking experts to use a rating scale for assessment. While this method does not account for the concordance among experts, it also introduces the possibility of bias, especially if the experts are aware of the responses provided by the LLMs. It is vital to note that GPT-4's responses, especially in complex cases like this study, can vary, impacting reproducibility [24]. Thus, only the program's first response was evaluated. It should also be noted that generative models can 'hallucinate' findings [25]. In our experiments, we provided GPT-4 with an explicit prompt and asked it to align responses with a clearly defined template, which might be less prone to hallucinationsthan 'open' questions.

Our study contributes to the ongoing discourse on the role of artificial intelligence in healthcare, particularly in the realm of medical diagnostics and risk stratification. The utilization of Large Language Models like ChatGPT-3.5 and ChatGPT-4, forms the backdrop of our exploration into innovative approaches for enhancing risk assessment tools. The potential of these models to aid in diagnosing NAFLD and predicting advanced fibrosis remains a subject of considerable interest, warranting further exploration and validation in future studies.

Finally, this study has a few limitations. First, the unavailability of API for GPT-4 and Bard limited our analysis to a subset of 9,254 participants. However, this result was sufficient to validate the findings of this study. Second, the inner workings of GPT-4 remain challenging because the model and the code of ChatGPT have not been fully disclosed, and because of the complex structure of LLM, fully explaining the working principle becomes difficult. Third, the performance of GPT-4 is yet to be extensively validated for various medical conditions, necessitating additional research to generalize our findings to other conditions such as diabetes or cancer. Studies are required to optimize the performance of GPT-4 through finetuning and prompt engineering of specific tasks. To fully confirm its effectiveness and practicality in real-world NAFLD diagnosis, further in-depth investigation and rigorous clinical.

VI. CONCLUSION

In this study, we validated the predictive performance of LLM in a prospective hospital-based cohort. The application of LLM-that incorporates various patient-level risk factors for NAFLD, such as smoking history, treatment, and laboratory factors- showed high discrimination (AUC of 0.831) as well as good calibration and risk stratification ability. Our research contributes to the evolving understanding of NAFLD's impact on diabetes, emphasizing the need for personalized risk assessment tools. The integration of advanced diagnostic calculations and consideration of glycemic status and fibrosis risk provide a more comprehensive evaluation of NAFLD severity and associated risks. As the field progresses, the application of AI models holds promise for refining risk prediction, necessitating further exploration and validation in diverse clinical settings. These findings underscore the imperative for continued research into innovative approaches to enhance patient care in the face ofthe growing NAFLD epidemic.

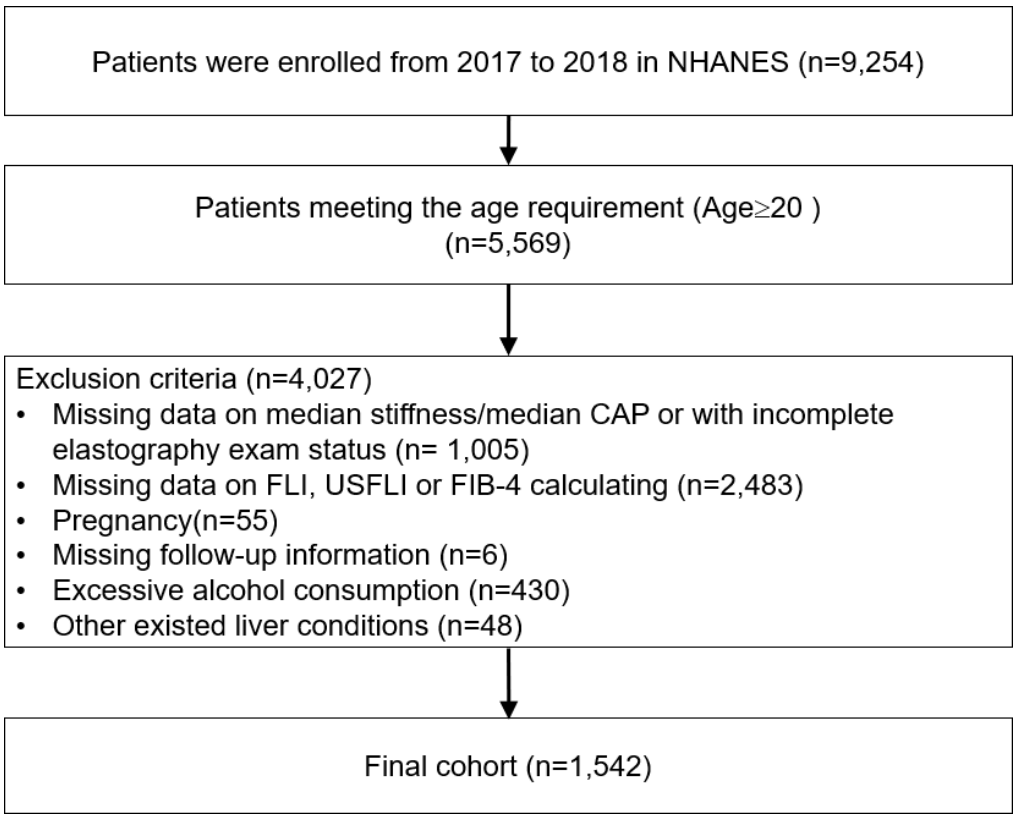

Figure1: The patient selection process

e-ISSN: 2583-553X
Peer Reviewed Journal
https://doi.org/10.5281/zenodo.13805421

Estimate the risk (in percentages) of developing an NAFLD, the severity of fibrosis if NAFLD is present, CVD mortality, and all-cause mortality within 10 years for the person below. 57-year-old female, non-smoker, total cholesterol 194.6 mg/dL, HDL 58.6 mg/dL, LDL 119.0mg/dL, triglyceride 63.3 mg/dL, systolic blood pressure 137 mmHg, diastolic blood pressure 86 mmHg, BMI 20.72, fasting glucose 6 mmol/L, AST 25 U/L, ALT 25 U/L, CRP 0.43 mg/dL, PLT 180*x109unit

Please answer exactly in the format below, without blank lines, and no further information or answer is required.
Risk percentage=(in percentages, round to one decimal place)

After that, give a brief description of why you give such a ratio

Risk percentage = (in percentages, round to one decimal place)

Risk of developing NAFLD: 25.0%

To estimate the risk of developing Non-Alcoholic Fatty Liver Disease (NAFLD) and the severity of fibrosis if NAFLD is present, we must consider various factors:

1. **Age and Gender**: At 57 years, this female patient is at a moderate age-related risk for NAFLD. Gender-wise, women generally have a lower risk compared to men, but post-menopausal women might have a higher risk, possibly due to changes in hormone levels.

2. **Blood Lipid Levels**: Her total cholesterol is moderately high, but her HDL (good cholesterol) is within a healthy range, and her LDL (bad cholesterol) and triglycerides are not elevated. These factors are favourable and suggest a lower risk of developing NAFLD.

3. **Blood Pressure**: Her systolic blood pressure is slightly elevated, which can be a risk factor for NAFLD, but it's not in the high hypertension range.

4. **BMI**: With a BMI of 20.72, she is within the normal weight range, which significantly lowers her risk of NAFLD, as obesity is a major risk factor.

5. **Blood Glucose Levels**: A fasting glucose level of 6 mmol/L is at the high end of the normal range, indicating a potential risk for prediabetes. Insulin resistance is a key factor in NAFLD development.

6. **Liver Enzymes (AST, ALT)**: Normal levels of AST and ALT suggest healthy liver function and lower the likelihood of existing liver damage or NAFLD.

7. **CRP Level**: A CRP of 0.43 mg/dL, which is within the normal range, suggests low systemic inflammation, reducing the risk of NAFLD development.

8. **Platelet Count (PLT)** : A normal platelet count also lowers the risk of severe liver fibrosis.

Figure 2: Example of a ChatGPT prompt and response for risk stratification

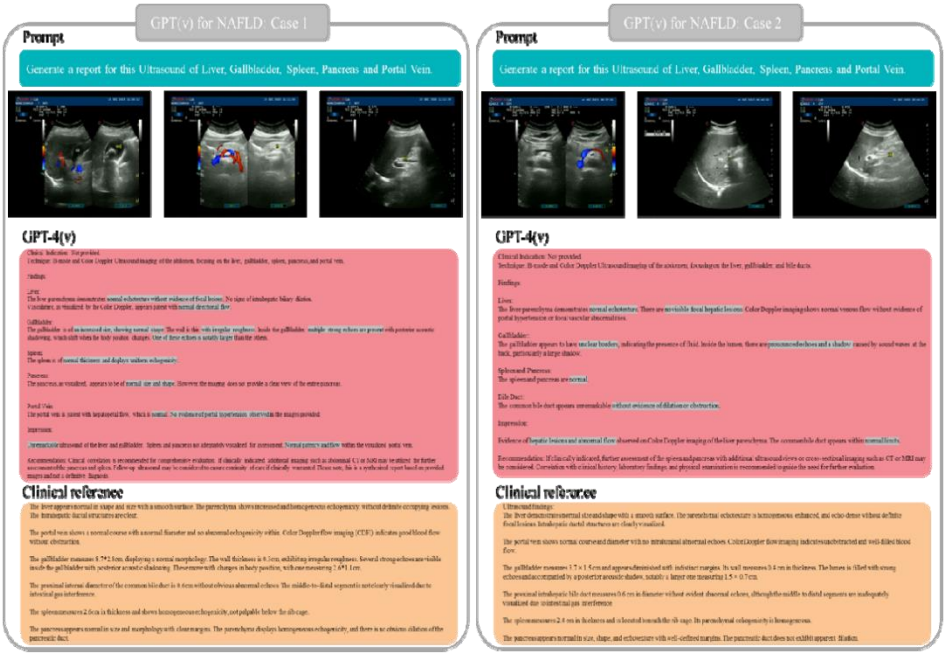

Figure 3: Example of a ChatGPT-4(V) prompt and response for risk stratification

e-ISSN: 2583-553X
Peer Reviewed Journal
https://doi.org/10.5281/zenodo.13805421

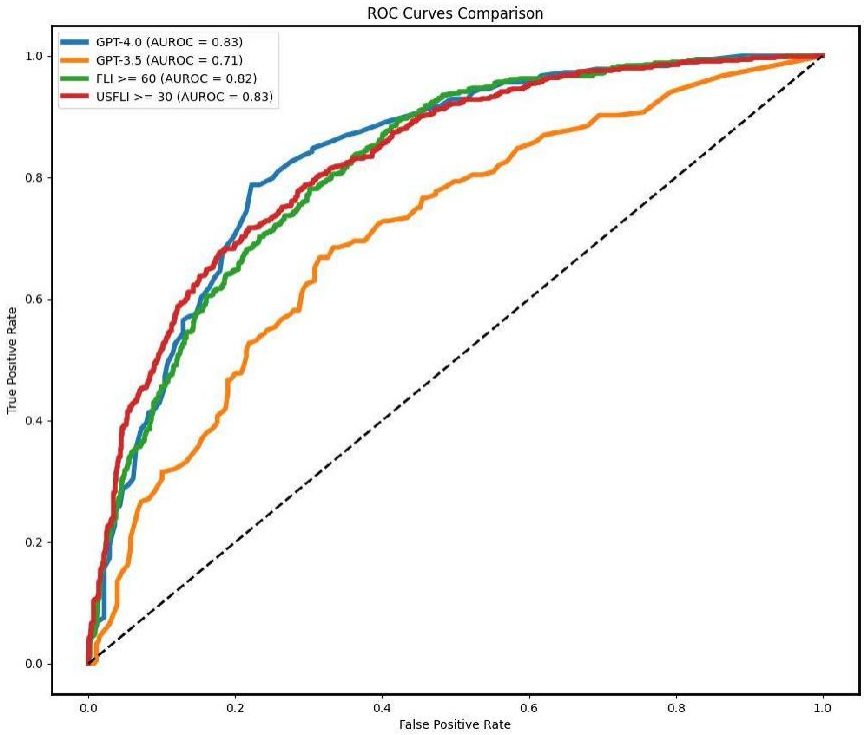

Figure 4: AUROC curve

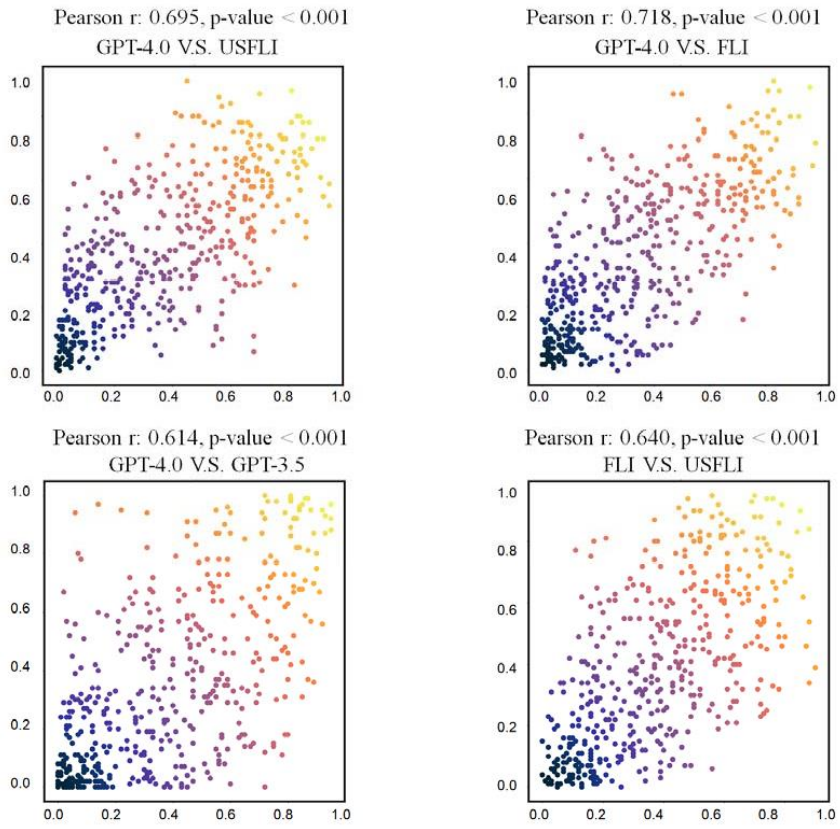

Figure 5: Scatterplots and Pearson correlation coefficient for various scoring methods

Model	AUROC	Brier	Sensitivity	Specificity	F1 Score
GPT-4.0	0.831 (0.796, 0.867)	0.167 (0.151, 0.183)	0.576 (0.506, 0.646)	0.853 (0.812, 0.893)	0.637 (0.578, 0.696)
GPT-3.5	0.707 (0.687, 0.727)	0.233 (0.213, 0.253)	0.556 (0.536, 0.576)	0.746 (0.726, 0.766)	0.571 (0.551, 0.591)
FLI>=60	0.817 (0.797, 0.837)	0.17 (0.160, 0.180)	0.66 (0.622, 0.697)	0.794 (0.768, 0.820)	0.66 (0.628, 0.692)
USFLI>=30	0.827 (0.807, 0.848)	0.172 (0.162,0.182)	0.464 (0.426, 0.503)	0.917 (0.899, 0.934)	0.58 (0.544, 0.616)
GPT-4.0 w/oDM	0.825 (0.784, 0.866)	0.165 (0.147, 0.182)	0.502 (0.415, 0.588)	0.875 (0.835, 0.916)	0.576 (0.498, 0.654)
GPT-3.5 w/o DM	0.707 (0.655, 0.758)	0.226 (0.198, 0.255)	0.524 (0.444, 0.604)	0.758 (0.706, 0.809)	0.536 (0.466, 0.607)
FLI w/o DM	0.817 (0.794, 0.839)	0.164 (0.153, 0.174)	0.638 (0.592, 0.685)	0.803 (0.778, 0.829)	0.628 (0.589, 0.666)
USFLI w/o DM	0.816 (0.793, 0.840)	0.168 (0.157, 0.178)	0.408 (0.360, 0.455)	0.929 (0.913, 0.946)	0.524 (0.479, 0.570)
GPT-4.0 with DM	0.749 (0.621, 0.878)	0.185 (0.144, 0.226)	0.84 (0.732, 0.947)	0.518 (0.286, 0.750)	0.808 (0.714, 0.901)
GPT-3.5 with DM	0.538 (0.359, 0.717)	0.289 (0.212, 0.366)	0.659 (0.500, 0.818)	0.524 (0.286, 0.762)	0.682 (0.559, 0.805)
FLI with DM	0.748 (0.675, 0.822)	0.209 (0.179, 0.239)	0.734 (0.662, 0.806)	0.628 (0.500, 0.755)	0.771 (0.712, 0.829)
USFL I with DM	0.758 (0.680, 0.837)	0.2 (0.169, 0.232)	0.647 (0.569, 0.725)	0.734 (0.623, 0.845)	0.735 (0.672, 0.798)

Table 1: Performance comparison of USFLI, FLI, and ChatGPT Risk Score

REFERENCES

1. Wang, Randi, & Morad Behandish. (2022). *Surrogate modeling for physical systems with preserved properties and adjustable tradeoffs*. arXiv preprint arXiv:2202.01139.
2. Li, Yinuo et al. (2023). Smart nanofiber mesh with locally sustained drug release enabled synergistic combination therapy for glioblastoma. *Nanomaterials (Basel, Switzerland), 13*(3), 414. doi:10.3390/nano13030414.
3. He, J., Li, F., Hu, X., Li, J., Nian, Y., Wang, J., ... & Tao, C. (2022, June). Chemical-protein relation extraction with pre-trained prompt tuning. in *IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pp. 608-609. IEEE.
4. Xiang, A., Qi, Z., Wang, H., Yang, Q., & Ma, D. (2024). *A multimodal fusion network for student emotion recognition based on transformer and tensor product*. arXiv preprint arXiv:2403.08511.
5. Wang, Yang, Yojiro Mori, & Hiroshi Hasegawa. (2021). Dynamic routing and spectrum allocation based on actor-critic learning for multi-fiber elastic optical networks. *Photonics in Switching and Computing*. Optica Publishing Group, 2021, pp. W1B.3. doi10.1364/PSC.2021.W1B.3.
6. Shimizu, Shosei et al. (2023). Boron neutron capture therapy for recurrent glioblastoma multiforme: Imaging

evaluation of a case with long-term local control and survival. *Cureus, 15*(1), e33898. doi:10.7759/cureus.33898.

7. Phan, T. M. N., Dao, C. T., Wu, C., Wang, J. Z., Liu, S., Ding, J. E., ... & Peng, W. C. (2024). *MEDFuse: Multimodal EHR data fusion with masked lab-test modeling and large language models*. arXiv preprint arXiv:2407.12309.
8. Wang, Randi, & Vadim Shapiro. (2019). Topological semantics for lumped parameter systems modeling. *Advanced Engineering Informatics, 42*, 100958.
9. Wang, Randi, Vadim Shapiro, & Morad Mehandish. (2024). Model consistency for mechanical design: bridging lumped and distributed parameter models with a priori guarantees. *Journal of Mechanical Design, 146*(5).
10. Lu, Q., Guo, X., Yang, H., Wu, Z., & Mao, C. (2024). Research on adaptive algorithm recommendation system based on parallel data mining platform. *Advances in Computer, Signals and Systems, 8*(5), 23-33.
11. Mo, Yuhong, et al. (2024). Large Language Model (LLM) AI text generation detection based on transformer deep learning algorithm. *International Journal of Engineering and Management Research, 14*(2), 154-159.
12. Ma, Boyuan, et al. (2021). Deep learning-based automatic inpainting for material microscopic images. *Journal of Microscopy, 281*(3), 177-189.
13. Ma, B., Ma, B., Gao, M., Wang, Z., Ban, X., Huang, H., & Wu, W. (2021). Deep learning-based automatic inpainting for material microscopic images. *Journal of Microscopy, 281*(3), 177-189.
14. Wu, Zekun, et al. (2022). Pipeline degradation evaluation based on distributed fiber sensors and convolutional neural networks (CNNs). *27th International Conference on Optical Fiber Sensors, Technical Digest Series*. Optica Publishing Group, Paper W4.41. doi:10.1364/OFS.2022.W4.41.
15. Li, Yinuo et al. (2023). A retrospective study of renal growth changes after proton beam therapy for pediatric malignant tumor. *Current Oncology (Toronto, Ont.), 30*(2), 1560-1570. doi:10.3390/curroncol30020120.
16. Li, Yinuo et al. (2022). Proton beam therapy for multifocal hepatocellular carcinoma (HCC) showing complete response in pathological anatomy after liver transplantation. *Cureusvol, 14*(6), e25744. doi:10.7759/cureus.25744.
17. Peng, Zhaoqiang, et al. (2019). Big data analytics on fiber-optical distributed acoustic sensing with rayleigh enhancements. *IEEE Photonics Conference (IPC)*, pp. 1-3. doi:10.1109/IPCon.2019.8908496.
18. Shimizu, Shosei et al. (2021). Proton beam therapy for a giant hepatic hemangioma: A case report and literature review. *Clinical and Translational Radiation Oncology, 27*, 152-156. doi:10.1016/j.ctro.2021.01.014.
19. Wang, Yu, et al. (2019). Particle filter vehicles tracking by fusing multiple features. *IEEE Access, 7*, 133694-133706.
20. Han, Yi, & Thomas CM Lee. (2024). Structural break detection in non-stationary network vector autoregression models. *IEEE Transactions on Network Science and Engineering*.
21. Yan, H., Wang, Z., Bo, S., Zhao, Y., Zhang, Y., & Lyu, R. (2024). *Research on image generation optimization based deep learning*.
22. Wang, Yang, et al. (2018). Optimizing multi-criteria k-shortest paths in graph by a natural routing genotype-based genetic algorithm. *13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 341-345. doi:10.1109/ICIEA.2018.8397739.
23. Mo, Yuhong, et al. (2024). Password complexity prediction based on roberta algorithm. *Applied Science and Engineering Journal for Advanced Research, 3*(3), 1-5.
24. Wang, Y., Ban, X., Wang, H., Li, X., Wang, Z., Wu, D., ... & Liu, S. (2019). Particle filter vehicles tracking by fusing multiple features. *IEEE Access, 7*, 133694-133706.
25. Wang, Qirui, et al. (2020). Pipeline defects detection and classification based on distributed fiber sensors and neural networks. *Optical Fiber Sensors Conference 2020 Special Edition, OSA Technical Digest, Optica Publishing Group*, Paper W2B.3. doi:10.1364/OFS.2020.W2B.3.
26. Restrepo, D., Nakayama, L. F., Dychiao, R. G., Wu, C., McCoy, L. G., Artiaga, J. C., ... & Celi, L. A. (2024, June). Seeing beyond borders: Evaluating LLMs in multilingual ophthalmological question answering. *IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pp. 565-566.
27. Yang, R. (2024). *CaseGPT: A case reasoning framework based on language models and retrieval-augmented generation*. arXiv preprint arXiv:2407.07913.
28. Nakamura, Masatoshi et al. (2024). A systematic review and meta-analysis of radiotherapy and particle beam therapy for skull base chondrosarcoma: TRP-chondrosarcoma. *Frontiers in Oncology, 14*, 1380716. doi:10.3389/fonc.2024.1380716.
29. He, J., Li, F., Li, J., Hu, X., Nian, Y., Xiang, Y., ... & Tao, C. (2024). Prompt tuning in biomedical relation extraction. *Journal of Healthcare Informatics Research, 8*(2), 206-224.
30. Nitta, Hazuki et al. (2024). An analysis of muscle growth after proton beam therapy for pediatric cancer. *Journal of Radiation Research, 65*(2), 251-255. doi:10.1093/jrr/rrad105.
31. Jin, Yonglong et al. (2023). Proton therapy (PT) combined with concurrent chemotherapy for locally advanced non-small cell lung cancer with negative driver genes. *Radiation Oncology (London, England), 18*(1), 189. doi:10.1186/s13014-023-02372-8.

32. Qi, Z., Ma, D., Xu, J., Xiang, A., & Qu, H. (2024). *Improved YOLOv5 based on attention mechanism and fasternet for foreign object detection on railway and airway tracks*. arXiv preprint arXiv:2403.08499.
33. Saito, Takashi et al. (2024). Systematic review and meta-analysis of particle beam therapy versus photon radiotherapy for skull base chordoma: TRP-chordoma 2024. *Cancersvol, 16*(14), 2569. doi:10.3390/cancers16142569.
34. Mo, Yuhong, et al. (2024). Make scale invariant feature transform "Fly" with CUDA. *International Journal of Engineering and Management Research, 14*(3), 38-45.
35. Li, Yinuo et al. (2024). Late changes in renal volume and function after proton beam therapy in pediatric and adult patients: children show significant renal atrophy but deterioration of renal function is minimal in the long-term in both groups. *Cancers, 16*(9), 1634. doi:10.3390/cancers16091634.
36. Huang, S., Yang, H., Yao, Y., Lin, X., & Tu, Y. (2024). *Deep adaptive interest network: personalized recommendation with context-aware learning*. arXiv preprint arXiv:2409.02425.
37. Wang, Yang, Yojiro Mori, & Hiroshi Hasegawa. (2020). Resource assignment based on core-state value evaluation to handle crosstalk and spectrum fragments in SDM elastic optical networks. *Opto-Electronics and Communications Conference (OECC)*, pp. 1-3. doi:10.1109/OECC48412.2020.9273621.
38. Xiang, A., Huang, B., Guo, X., Yang, H., & Zheng, T. (2024). *A neural matrix decomposition recommender system model based on the multimodal large language model*. arXiv preprint arXiv:2407.08942.
39. Niitsu, Hikaru et al. (2024). Tumor response on diagnostic imaging after proton beam therapy for hepatocellular carcinoma. *Cancers, 16*(2), 357. doi:10.3390/cancers16020357.
40. Kumada, Hiroaki et al. (2022). *Current development status of iBNCT001, demonstrator of a LINAC-based neutron source for BNCT*. doi:10.3233/JNR-220029.
41. Hu, X., Sun, Z., Nian, Y., Wang, Y., Dang, Y., Li, F., ... & Tao, C. (2024). Self-explainable graph neural network for alzheimer disease and related dementias risk prediction: algorithm development and validation study. *JMIR Aging, 7*(1), e54748.
42. Han, Yi, & Thomas Lee. (2024). *Change point detection in pairwise comparison data with covariates*. arXiv preprint arXiv:2408.13642.
43. Li, F., Rasmy, L., Xiang, Y., Feng, J., Abdelhameed, A., Hu, X., ... & Tao, C. (2024). Dynamic prognosis prediction for patients on DAPT after drug-eluting stent implantation: Model development and validation. *Journal of the American Heart Association, 13*(3), e029900.
44. Fan, X., Tao, C., & Zhao, J. (2024). Advanced stock price prediction with xLSTM-based models: Improving long-term forecasting. *Preprints*, 2024082109. https://doi.org/10.20944/preprints202408.2109.v1.
45. Dang, B., Ma, D., Li, S., Qi, Z., & Zhu, E. (07 2024). Deep learning-based snore sound analysis for the detection of night-time breathing disorders. *Applied and Computational Engineering, 76*, 109–114. doi:10.54254/2755-2721/76/20240574.
46. He, Shuyao, et al. (2024). Lidar and monocular sensor fusion depth estimation. *Applied Science and Engineering Journal for Advanced Research, 3*(3), 20-26.
47. Dai, Shuying, et al. (2024). The cloud-based design of unmanned constant temperature food delivery trolley in the context of artificial intelligence. *Journal of Computer Technology and Applied Mathematics, 1*(1), 6-12.
48. Li, S., Dong, X., Ma, D., Dang, B., Zang, H., & Gong, Y. (2024). Utilizing the LightGBM algorithm for operator user credit assessment research. *Applied and Computational Engineering, 75*(1), 36–47. doi:10.54254/2755-2721/75/20240503.
49. Zhan, Donglin, et al. (2019). *Adaptive transfer learning of multi-view time series classification*. arXiv preprint arXiv:1910.07632.
50. Ma, D., Li, S., Dang, B., Zang, H., & Dong, X. (2024). Fostc3net: A lightweight YOLOv5 based on the network structure optimization. *Journal of Physics: Conference Series, 2824*(1), 012004. doi:10.1088/1742-6596/2824/1/012004.
51. Song, Jintong, et al. (2024). A comprehensive evaluation and comparison of enhanced learning methods. *Academic Journal of Science and Technology, 10*(3), 167-171.
52. Dai, Shuying, et al. (2024). AI-based NLP section discusses the application and effect of bag-of-words models and TF-IDF in NLP tasks. *Journal of Artificial Intelligence General Science (JAIGS), 5*(1), 13-21.
53. Tang, X., Wang, Z., Cai, X., Su, H., & Wei, C. (2024). *Research on heterogeneous computation resource allocation based on data-driven method*. arXiv preprint arXiv:2408.05671.
54. Li, W., Li, H., Gong, A., Ou, Y., & Li, M. (2018, August). An intelligent electronic lock for remote-control system based on the internet of things. in *Journal of Physics: Conference Series, 1069*(1), 012134. IOP Publishing.
55. Li, Shaojie, Yuhong Mo, & Zhenglin Li. (2022). Automated pneumonia detection in chest x-ray images using deep learning model. *Innovations in Applied Engineering and Technology*, 1-6.
56. Li, Keqin, et al. (2024). Utilizing deep learning to optimize software development processes. *Journal of Computer*

Technology and Applied Mathematics, 1(1), 70-76.
57. Wu, Z., Wang, X., Huang, S., Yang, H., & Ma, D. (2024). Research on prediction recommendation system based on improved markov model. *Advances in Computer, Signals and Systems, 8*(5), 87-97.
58. Tan, C., Wang, C., Lin, Z., He, S., & Li, C. (2024). Editable neural radiance fields convert 2d to 3d furniture texture. *International Journal of Engineering and Management Research, 14*(3), 62-65.
59. Li, Zhenglin, et al. (2023). Stock market analysis and prediction using LSTM: A case study on technology stocks. *Innovations in Applied Engineering and Technology*, 1-6.
60. Liu, Tianrui, et al. (2024). Spam detection and classification based on distilbert deep learning algorithm. *Applied Science and Engineering Journal for Advanced Research, 3*(3), 6-10.
61. Liu, Jihang, et al. (2024). Unraveling large language models: From evolution to ethical implications-introduction to large language models. *World Scientific Research Journal, 10*(5), 97-102.
62. Zhu, Armando, et al. (2024). Cross-task multi-branch vision transformer for facial expression and mask wearing classification. *Journal of Computer Technology and Applied Mathematics, 1*(1), 46-53.
63. Li, Keqin, et al. (2024). The application of augmented reality (AR) in remote work and education. *Journal of Computer Technology and Applied Mathematics, 1*(1), 33-39.
64. Hong, Bo, et al. (2024). The application of artificial intelligence technology in assembly techniques within the industrial sector. *Journal of Artificial Intelligence General Science (JAIGS), 5*(1), 1-12.
65. Zhao, Peng, et al. (2024). Task allocation planning based on hierarchical task network for national economic mobilization. *Journal of Artificial Intelligence General Science (JAIGS), 5*(1), 22-31.
66. Liu, Minghao, Qing Wang, & Wei Lu. (2017). Peridynamic simulation of brittle-ice crushed by a vertical structure. *International Journal of Naval Architecture and Ocean Engineering, 9*(2), 209-218.
67. Liu, Minghao, & Jay Oswald. (2019). Coarse–grained molecular modeling of the microphase structure of polyurea elastomer. *Polymer, 176*, 1-10.
68. Zan, Yingfei, et al. (2014). Research on real-time simulation system of ship motion based on simulink. *The Open Mechanical Engineering Journal, 8*(1).
69. Cui, Xiongwei, et al. (2017). A hybrid wavelet-based adaptive immersed boundary finite-difference lattice Boltzmann method for two-dimensional fluid–structure interaction. *Journal of Computational Physics, 333*, 24-48.
70. Cui, Xiongwei, et al. (2018). A coupled Volume Penalization-Thermal Lattice Boltzmann method for thermal flows. *International Journal of Heat and Mass Transfer, 127*, 253-266.
71. Cui, Xiongwei, et al. (2022). A coupled two-relaxation-time lattice boltzmann-volume penalization method for flows past obstacles. *Mathematics and Computers in Simulation, 198*, 85-105.
72. Guo, Kai, Xiongwei Cui, & Minghao Liu. (2018). A coupled lattice boltzmann-volume penalization for flows past fixed solid obstacles with local mesh refinement. *Mathematical Problems in Engineering, 2018*(1), 6732082.
73. Liu, Minghao, Jianlan Ye, & Jay Oswald. (2023). Coarse-grained molecular simulation of the role of curing rates on the structure and strength of polyurea. *Computational Materials Science, 230*, 112428.
74. Phan, Thao Minh Nguyen, et al. (2024). *MEDFuse: Multimodal EHR data fusion with masked lab-test modeling and large language models*. arXiv preprint arXiv:2407.12309.
75. Ding, Jun-En, et al. (2024). Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records. *Scientific Reports, 14*(1), 20774.