Establishment of a Knowledge-and-Data-Driven Artificial Intelligence System with Robustness and Interpretability in Laboratory Medicine

Beilei Wang, Jie Jing, Xiaochun Huang, Cheng Hua, Qin Qin, Yin Jia, Zhiyong Wang, Lei Jiang, Bai Gao, Lele Wu, Xianfei Zeng, Fubo Wang, Chuanbin Mao, and Shanrong Liu

1. Introduction

Laboratory medicine plays an important role in clinical diagnosis. However, no laboratory-based artificial intelligence (AI) diagnostic system has been applied in current clinical practice due to the lack of robustness and interpretability. Although many attempts have been made, it is still difficult for doctors to adopt the existing machine learning (ML) patterns in interpreting laboratory (lab) big data. Here, a knowledge-and-data-driven laboratory diagnostic system is developed, termed AI-based Lab tEst tO diagNosis (AI LEON), by integrating an innovative knowledge graph analysis framework and "mixed XGBoost and Genetic Algorithm (MiXG)" technique to simulate the doctor’s laboratory-based diagnosis. To establish AI LEON, we included 89,116,949 laboratory data and 10,423,581 diagnosis data points from 730,113 participants. Among them, 686,626 participants were recruited for training and validating purposes with the remaining for testing purposes. AI LEON automatically identified and analyzed 2071 lab indexes, resulting in multiple disease recommendations that involved 441 common diseases in ten organ systems. AI LEON exhibited outstanding transparency and interpretability in three universal clinical application scenarios and outperformed human physicians in interpreting lab reports. AI LEON is an advanced intelligent system that enables a comprehensive interpretation of lab big data, which substantially improves the clinical diagnosis.
extremely limited due to the breadth and complexity of capturing all the medical knowledge.[9] Recently, deep learning (DL) models have made significant progress in automatically extracting rules from large amounts of data, rather than explicitly specified by experts.[10] Due to high performance, these data-driven AI systems have made many breakthroughs, such as sound recognition and image recognition (radiology, pathology, dermatology, etc.).[12,10–12] However, data availability and reliability is a foundational challenge as the quality of the output is highly dependent on the quality of the input.[13] In addition, the accurate and scalable data interpretation is another major challenge due to the US Food and Drug Administration’s “black box” warning on DL models.[14] In medical diagnosis, we need not only to learn from prior data to extract and generalize knowledge[15] but also to clarify the complex inter-relationships and potential explanatory factors of the knowledge.[16] Therefore, the key to reaching truly usable AI is to develop a transparent, interpretable, trustworthy, and robust medical diagnostic system through the integration of knowledge-driven and data-driven models.

As we all know, laboratory tests have been widely used to exclude, confirm, classify, or monitor various diseases to assist doctors in their clinical decision. For example, more than 41% of emergency patients require a laboratory test. Many diseases, such as diabetes, cancers, thrombotic disorders, and genetic diseases, cannot be diagnosed without laboratory testing.[15,16] More importantly, laboratory data are needed to support evidence-based medical practice and the development of guidelines for clinical diagnosis and treatment of many diseases. For example, out of 1230 evidence-based clinical practice guidelines for 23 major diseases, 37% involved laboratory tests.[17] Collectively, laboratory medicine can provide a lot of factual, explicit, and tacit knowledge for clinical practice and is therefore very suitable for establishing knowledge-based interpretable AI diagnostic systems. In contrast, compared with imaging data that require manual annotations and careful quality control, laboratory data provide more advantages. First, in recent two decades, automation has dramatically increased the capacity and improved the standardization of clinical laboratories, thus generating vast structured or semistructured data.[17,18] Second, over the last 30 years, much progress has been made to improve the standardization and harmonization of laboratory methods and results.[16,19] For example, the International Consortium for Harmonization of Clinical Laboratory Results (ICHCLR) has paid attention to the unification of laboratory results.[19] Therefore, a large volume of high-quality laboratory data is also very suitable for establishing data-driven diagnostic systems.

In laboratory-based diagnosis, the doctors usually formulate a checklist for laboratory tests based on chief complaints. By evaluating abnormal laboratory results, the doctors compile a list of possible underlying diseases on the basis of their knowledge and experience. For those complicated diseases, the doctors may further add laboratory tests to increase the accuracy of the diagnosis. Then, the most severe diseases (such as malignant tumors) and related diseases (such as hematuria) are ranked first, followed by other minor diseases (such as malnutrition). Therefore, laboratory-based diagnosis is not only a classification process but also a complex ranking one. To mimic the decision-making processes of human doctors, we need to develop a novel AI diagnostic system that not only calculates the probabilities of different diseases but also automatically ranks them according to their severity and relevance. Moreover, to better assist decision-making, such a system should also be transparent, interpretable, and comprehensible for human doctors.

In this study, we tried to develop a novel knowledge-and-data-driven AI diagnostic system, termed “AI-based Lab Test to diagnose” (AI LEON). At the technical level, we successfully integrated knowledge-driven and data-driven models by combining a comprehensive knowledge graph analysis framework and a state-of-the-art “mixed XGBoost and Genetic Algorithm (MiXG)” technique. At the application level, we implemented multiple disease recommendations in clinical practice and simulated the doctor’s laboratory-based disease reasoning process through the effective integration of medical knowledge, big medical data, and real-world diseases.

2. Results

2.1. Characteristics of the Participants

Between Jan 1, 2010, and Oct 1, 2020, 1548185 participants who had laboratory reports within 21 days prior to diagnosis were included in the study. After disease annotation, 818072 participants were excluded because of undetermined diagnoses by laboratory medicine. Finally, 730113 eligible participants with 1604596 visits, 2071 laboratory indexes with 89116949 data, and 441 common diseases with 10423581 diagnostic data were included and analyzed in the study. All the laboratory indexes were categorized into 12 sections: routine tests, clinical biochemistry, hematology, coagulation, endocrinology, microbiology, immunology tests, body fluid analysis, antibiotic susceptibilities, flow cytometry cell markers, cytology, and molecular diagnostics.
For the retrospective cohort, 686,626 participants were randomly divided into a training set (509,841 participants and 65,034,574 laboratory data) and a validation set (235,074 participants and 16,268,666 laboratory data). These data were explored to develop and validate AI LEON, which was eventually tested in a large cohort of 69,101 participants with 7,813,709 laboratory data and 370 common diseases. A summary of the workflow is provided in Figure 1. The clinical characteristics of the eligible participants are detailed in Table 1. The participants had a median age of 56 years (interquartile range from 38 to 67 years) with 54.58% male. All the diseases were categorized into ten organ systems: infection, respiratory system, hematological system, gastrointestinal system, urinary system, endocrine system, circulatory system, immune system, reproductive system, and nervous system (Figure 2). The most frequently encountered diseases of each system are listed in Table 1.

2.2. AI LEON to Simulate Physicians’ Diagnosis

To mimic human physicians’ clinical diagnostic process in real time, AI LEON was established using two core models, the knowledge-driven model (Figure 3a) and the data-driven model (Figure 3b). In this system, the knowledge-driven models could diagnose 273 commonly encountered diseases, whereas the data-driven models could predict the probability of the remaining 168 diseases. The former first provided evidence of the likelihood of the diseases and the latter would be triggered to further calculate the probability of the diseases once the former failed to diagnose them directly. The workflow of AI LEON is depicted in Figure 3c, demonstrating how it works in the clinic. First, all the laboratory data from the participants are systematically evaluated by the system. Then, the possible organ systems are recommended, followed by listing the specific diseases according to the different risk levels. Ultimately, the platform is expected to assist doctors in making accurate and timely clinical interventions.

2.3. AI LEON Showing High Accuracy for Disease Diagnosis

AI LEON was independently tested in the prospective test cohort. The diagnostic performance of the 370 frequently encountered diseases in ten organ systems was assessed by the system. Among them, 203 diseases were directly diagnosed by the knowledge-driven models of the system. The other 167 possible diseases were first evaluated by the knowledge-driven models and then the probabilities of the potential diseases by the data-driven models were calculated.

For the first level of diagnosis, a high level of accuracy was achieved to predict the involved organ systems; the overall mean average precision (mAP) was 87.74%, ranging from 81.43% for the circulatory system to 96.96% for infection (See Table 2). For the final level of diagnosis, the system showed a solid performance in diagnosing diseases with a median mAP of 84.40%. The diagnostic performance was also evaluated in organ-based subcohorts of patients. Under a 100% recall rate across all subcohorts, a high level of mAP was achieved, ranging from 74.54% for the nervous system to 91.72% for infection (See Table 2).

In addition, mAP of top n for broad organ systems and the corresponding recalls were evaluated (Table 3, Figure 4a). For instance, if the top five organ systems were recommended, a mAP of 89.83% and a recall rate of 92.80% were achieved,
Table 1. Baseline characteristics.

| Organ systems       | Incidence | Training cohort | Validation cohort | Testing cohort | Total       |
|---------------------|-----------|-----------------|-------------------|---------------|-------------|
| Infection           | 28.329%   | 278 832 (54.69%)| 126 540 (53.83%)  | 39 132 (56.63%)| 398 496 (54.58%)|
| Viral infection     |           |                 |                   |               |             |
| Herpes simplex      |           |                 |                   |               |             |
| Tuberculosis        |           |                 |                   |               |             |
| Parasitic disease   |           |                 |                   |               |             |
| Respiratory system  | 21.037%   |                 |                   |               |             |
| Upper respiratory tract infection | | 699 447 | 174 798 | 55 287 | 929 532 |
| Bronchitis          |           |                 |                   |               |             |
| Lung cancer         |           |                 |                   |               |             |
| Community acquired pneumonia | | 4594 | 1114 | 1251 | 6959 |
| Pulmonary embolism  |           |                 |                   |               |             |
| Hematological system| 19.128%   |                 |                   |               |             |
| Small cell hypochromic anemia | | 41 714 | 10 430 | 5767 | 57 911 |
| Lymphoma            |           |                 |                   |               |             |
| Acute leukemia      |           |                 |                   |               |             |
| Multiple myeloma    |           |                 |                   |               |             |
| Aplastic anemia     |           |                 |                   |               |             |
| Hematological system| 13.888%   |                 |                   |               |             |
| Liver cirrhosis     |           |                 |                   |               |             |
| Pancreatic cancer   |           |                 |                   |               |             |
| Chronic hepatitis B |           |                 |                   |               |             |
| Pancreatitis        |           |                 |                   |               |             |
| Colon cancer        |           |                 |                   |               |             |
| Urinary system      | 5.935%    |                 |                   |               |             |
| Renal insufficiency |           |                 |                   |               |             |
| Urinary system infection | | 44 852 | 11 120 | 5850 | 61 822 |
| Urinary system stones |        |                 |                   |               |             |
| Glomerulus nephritis|           |                 |                   |               |             |
| Urinary Tract Tumors|          |                 |                   |               |             |
| Endocrine system    | 5.832%    |                 |                   |               |             |
| Diabetes            |           |                 |                   |               |             |
| Hyperlipidemia      |           |                 |                   |               |             |
| Hyperthyroidism     |           |                 |                   |               |             |
| Hypothyroidism      |           |                 |                   |               |             |
| Acidosis            |           |                 |                   |               |             |
| Circulatory system  | 3.904%    |                 |                   |               |             |
| Atherosclerosis     |           |                 |                   |               |             |
| Coronary heart disease |       |                 |                   |               |             |
| Hypertension        |           |                 |                   |               |             |
| Myocardial infarction|          |                 |                   |               |             |
| Heart failure       |           |                 |                   |               |             |
| Total number of patients |       | 509 841 | 235 074 | 69 101 | 730 113 |
| Median age at visit (years) |   | 56 | 56 | 56 | 56 |
| Number of clinical samples | | 1 188 477 | 296 740 | 119 379 | 1 604 596 |
| Laboratory data     | 65 034 574 | 16 268 666 | 7 813 709 | 89 116 949 |
| Diagnose data       | 7 600 838 | 1 897 653 | 925 090 | 10 423 581 |
Table 1. Continued.

| System                     | Training cohort | Validation cohort | Testing cohort | Total |
|----------------------------|-----------------|-------------------|----------------|-------|
| Immune system              | 1.379%          |                   |                |       |
| Rheumatoid arthritis       | 16748           | 4093              | 1941           | 22782 |
| Systemic lupus erythematosus | 6210            | 1574              | 628            | 8412  |
| Gout                       | 5790            | 1422              | 423            | 7635  |
| Ankylosing spondylitis     | 4741            | 1147              | 390            | 6278  |
| IgA nephropathy            | 1880            | 477               | 412            | 2769  |
| Reproductive system        | 0.560%          |                   |                |       |
| Prostate cancer            | 8629            | 2187              | 1519           | 12335 |
| Enlarged or hyperplastic prostate | 8254  | 2148              | 845            | 11247 |
| Breast cancer              | 4975            | 1165              | 519            | 6659  |
| Ovarian cyst               | 1166            | 309               | 122            | 1597  |
| Nervous system             | 0.007%          |                   |                |       |
| Cerebral infarction        | 3,823           | 922               | 492            | 5237  |
| Brain tumor                | 123             | 31                | 21             | 175   |

Figure 2. Hierarchy of the diagnostic framework of AI LEON. Using an organ-based approach, the diagnoses are divided first into ten broad organ systems and then subsequently into specific disease groups under each organ system. Abbreviations: AI LEON, Artificial Intelligence-based Lab Test to Diagnose.
Figure 3. Overview of AI LEON. a) The basic working mechanism of the knowledge-driven models in AI LEON. The output results of the knowledge-driven model are divided into two categories, including definite diagnosis and suspicious diagnosis. The former shows very strong confidence in disease diagnosis (e.g., Yes/No) based on a combination of lab results. The latter shows relatively low confidence in some diseases (e.g., Sus). b) The basic working mechanism of the data-driven models in AI LEON. Suspicious diagnosis suggested by the knowledge-driven model triggers the corresponding data-driven model, which further calculates the probability and relevance of the diseases. c) The workflow of AI LEON. The platform systematically evaluates the overall results of laboratory tests from the population. Then, the possible organ systems are recommended, followed by listing the specific diseases according to the different risk levels of the diseases. Ultimately, the AI LEON platform is expected to assist the doctors in making timely clinical interventions. Abbreviations: AI LEON, Artificial Intelligence-based Lab tEst tO diagnosis; Sus, Suspicious diagnosis.
respectively. If top 10 organ systems were recommended, mAP was 87.74%, and the recall rate was increased to 99.96%. Further, mAP of top \( n \) for all diseases and the corresponding recalls were also evaluated (Table 3, Figure 4b). High accuracy of mAP of 91.95% was achieved when the top five diseases were diagnosed by AI LEON with a recall rate of 66.99%. When the top 10 diseases were diagnosed, the recall rate was increased to 85.41% with mAP reaching 88.85%.

Similar results were observed in ten organ-based subcohorts of patients (Table 4, Figure 4c–l). For example, AI LEON has the highest accuracy of mAP in diagnosing a subcohort of patients with infectious diseases. When five infectious diseases were recommended, mAP was 95.67%, and the recall rate was 69.81%; when 10 infectious diseases were recommended, the mAP was slightly reduced to 94.26%, and the recall rate was increased to 87.06%. These results indicated the missing of very few diseases and a very high degree of accuracy when ten diseases were diagnosed by AI LEON.

2.4. AI LEON Showing Good Clinical Interpretability

To further illustrate how AI LEON provides the logical reasoning process of disease diagnosis, we envisioned three clinical application scenarios. The first scenario was named the “gold standard-based diagnosis” using knowledge-driven models based on the guidelines and expert consensus (Figure 5a). Taking COVID-19 as an example, the disease can be confirmed by the COVID-19 RNA test report, when at least two genes (ORF1ab, N and E) were positive.

The second scenario was called “specific diagnosis” using etiology-related knowledge-driven models (Figure 5b). The first knowledge-driven model is triggered to make a preliminary diagnosis, and then the second or more knowledge-driven models are triggered to give the definitive diagnosis. Take renal anemia as an example; the definitive diagnosis is finally made by integrating the anemia diagnostic formula and chronic kidney disease (CKD) formula.

The third scenario was termed “differential diagnosis” by combining knowledge-driven models with data-driven models. As shown in Figure 5c, an abnormal laboratory parameter usually represents several potential diseases predicted by the knowledge-driven model. Subsequently, the corresponding data-driven models automatically evaluate the probability of these diseases using all the available laboratory parameters. For example, the knowledge-driven model predicts that the increase in CA19-9 may be accompanied with various diseases, such as pancreatic cancer, pancreatitis, ovarian cyst, etc. Then, the corresponding data-driven models will be triggered to calculate the probabilistic risks of these possible diseases. Finally, an ordered list of diagnoses is ranked in the order of probability from the highest to the lowest.

2.5. AI LEON Outperforms Human Physicians in Laboratory-Based Diagnosis

We also compared the performance of diagnosis between AI LEON and 15 certified physicians in interpreting lab reports of 100 cases from an independent test cohort. The reader study consists of a blank-filling study and a true/false study. In the blank-filling study, the accuracy of AI LEON was comparable

| Table 2. Performance of artificial intelligence-based Lab tEst tOdiagnosis (AI LEON) in the prospective test cohort. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Overall | Subcohorts |
| | Infection system | Respiratory system | Hematological system | Circulatory system | Immune system | Gastrointestinal system | Endocrine system | Nervous system | Urinary system | Reproductive system |
| Organ system | mAP | 87.74% | 96.96% | 92.57% | 83.01% | 81.43% | 91.68% | 83.89% | 88.01% | 92.97% | 88.86% | 84.99% |
| Recall | 100% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |

| Table 3. mAP of top \( n \) and the corresponding recalls. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TOP 1 | TOP 2 | TOP 3 | TOP 4 | TOP 5 | TOP 6 | TOP 7 | TOP 8 | TOP 9 | TOP 10 |
| Organ system | mAP | 89.91% | 92.16% | 91.95% | 90.92% | 89.83% | 88.98% | 88.40% | 88.01% | 87.81% | 87.75% |
| Recall | 42.71% | 67.74% | 80.53% | 87.87% | 92.80% | 95.99% | 98.08% | 99.28% | 99.81% | 99.96% |

| Diseases | mAP | 91.44% | 92.87% | 93.00% | 92.61% | 91.95% | 91.25% | 90.57% | 89.93% | 89.35% | 88.85% |
| Recall | 20.26% | 37.44% | 50.45% | 59.97% | 66.99% | 72.34% | 76.64% | 80.12% | 83.04% | 85.41% |
with that of each general practitioner (GP) group (Figure 6a and Table 5, AI LEON: 0.791 vs GP groups: 0.637, 0.487, 0.499, respectively). The AI LEON’s mAP was much better than that of all the three GP groups in both the overall patients (Figure 6a and Table 5, AI LEON: 0.862 vs GP groups: 0.183, 0.183, 0.176, respectively) and organ-based sub-group of patients (Figure 6b–k and Table 6). AI LEON also achieved a higher recall rate of diagnoses than each GP group (Figure 6a and Table 5, AI LEON: 0.919 vs GP groups: 0.165, 0.168, 0.162, respectively), indicating that our system could give a more comprehensive diagnosis and avoid the rate of missed diagnosis. Similarly, our results demonstrated that AI LEON displayed better diagnostic performance than human physicians in the true–false study (Figure 6l–v and Table 6). Take together, our results indicated that AI LEON may be used to improve the clinical diagnosis of human physicians.

3. Discussion

Here, we presented a novel interpretable knowledge-and-data-driven AI system (AI LEON) to predict the diagnoses by
It is notable that the diagnosis process is more of a ranking process than a simple binary diagnosis. Therefore, we proposed an innovative MiXG algorithm that first calculated the probabilities of different diseases from the XGBoost classifiers and then adopted the mAP metric from the GA models to ensure that the most relevant and severe diagnostic result was ranked first. The mAP value of AI LEON reached 89.83% and 91.95% for predicting the top five organ systems and the top five specific diseases, respectively (Table 3), clearly proving the high accuracy and completeness of our system. Clinically, it remains a daunting challenge to identify all the possible underlying diseases, especially when evaluating the diseases on the basis of laboratory tests alone. Using routine laboratory tests as input, our platform succeeded in making reasonably accurate diagnosis predictions in both organ systems as well as specific diseases. Then, we further performed a human–machine competition and confirmed that AI LEON outperformed human physicians in laboratory-based diagnosis (Table 5 and 6). Based on these findings, we believe that our system may help physicians improve their diagnostic accuracy.

Many automated machine learning (ML) diagnostic models have two advantages. First, it exhibits excellent robustness and accuracy in laboratory disease diagnosis. The performance of ML models is known to continue to improve as the amount of input data increases. Herein, we used the massive volume of data, including 89,116,949 lab data and 10,423,581 diagnostic results, to develop and test the models, significantly improving the robustness of the diagnostic system. In addition, we used the knowledge graph to link the lab data with the diagnostic data, to develop and test the models, significantly improving the robustness of the diagnostic system. In addition, we used the knowledge graph to link the lab data with the diagnostic data, to develop and test the models, significantly improving the robustness of the diagnostic system. In addition, we used the knowledge graph to link the lab data with the diagnostic data, to develop and test the models, significantly improving the robustness of the diagnostic system. In addition, we used the knowledge graph to link the lab data with the diagnostic data, to develop and test the models, significantly improving the robustness of the diagnostic system.

| Table 4. mAP of top n and the corresponding recalls in subgroups. |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                 | TOP1             | TOP2             | TOP3             | TOP4             | TOP5             | TOP6             | TOP7             | TOP8             | TOP9             | TOP10            |
| Infection       | mAP 91.72%       | 95.45%           | 96.66%           | 96.48%           | 96.12%           | 95.67%           | 95.30%           | 94.99%           | 94.72%           | 94.48%           | 94.26%           |
| Recall          | 100.00%          | 21.23%           | 39.75%           | 52.93%           | 62.54%           | 69.81%           | 75.19%           | 79.27%           | 82.44%           | 84.99%           | 87.06%           |
| Respiratory     | mAP 88.19%       | 93.18%           | 94.31%           | 94.99%           | 94.19%           | 93.23%           | 92.55%           | 92.03%           | 91.64%           | 91.28%           | 91.01%           |
| System          | Recall 100.00%    | 19.72%           | 38.25%           | 53.89%           | 63.51%           | 70.73%           | 75.87%           | 79.73%           | 82.54%           | 84.88%           | 86.71%           |
| Endocrine       | mAP 85.64%       | 93.90%           | 95.34%           | 95.18%           | 94.58%           | 93.74%           | 92.81%           | 91.95%           | 91.20%           | 90.56%           | 89.98%           |
| Recall          | 100.00%          | 18.00%           | 32.50%           | 44.27%           | 53.57%           | 60.56%           | 66.24%           | 71.02%           | 74.98%           | 78.35%           | 81.23%           |
| Urinary system  | mAP 85.40%       | 95.52%           | 96.35%           | 96.19%           | 95.63%           | 94.85%           | 94.05%           | 93.19%           | 92.39%           | 91.66%           | 91.06%           |
| Recall          | 100.00%          | 12.79%           | 24.61%           | 35.27%           | 44.48%           | 52.24%           | 58.82%           | 64.59%           | 69.57%           | 73.89%           | 77.44%           |
| Immune system   | mAP 83.48%       | 92.22%           | 93.66%           | 93.30%           | 92.55%           | 91.35%           | 90.31%           | 89.56%           | 88.88%           | 87.94%           | 87.94%           |
| Recall          | 100.00%          | 13.98%           | 25.87%           | 36.28%           | 45.11%           | 52.58%           | 58.85%           | 63.94%           | 68.34%           | 72.12%           | 75.40%           |
| Gastrointestinal system | mAP 78.23% | 91.00%           | 92.29%           | 92.26%           | 91.73%           | 90.82%           | 89.91%           | 88.95%           | 88.09%           | 87.24%           | 86.45%           |
| Recall          | 100.00%          | 14.96%           | 27.84%           | 38.60%           | 47.34%           | 54.48%           | 60.26%           | 65.23%           | 69.44%           | 73.14%           | 76.36%           |
| Hematological   | mAP 77.18%       | 92.00%           | 92.20%           | 92.20%           | 92.14%           | 91.85%           | 91.33%           | 90.73%           | 90.01%           | 89.24%           | 88.56%           |
| System          | Recall 100.00%    | 15.29%           | 29.59%           | 41.58%           | 50.48%           | 56.83%           | 61.68%           | 65.90%           | 69.49%           | 72.66%           | 75.35%           |
| Circulatory     | mAP 76.06%       | 91.82%           | 92.71%           | 92.82%           | 92.45%           | 91.68%           | 90.64%           | 89.52%           | 88.36%           | 87.31%           | 86.41%           |
| Recall          | 100.00%          | 12.86%           | 24.52%           | 34.33%           | 42.44%           | 49.20%           | 55.03%           | 60.23%           | 64.74%           | 68.87%           | 72.38%           |
| Nervous system  | mAP 74.54%       | 88.10%           | 88.10%           | 88.10%           | 90.36%           | 89.00%           | 86.77%           | 86.61%           | 85.92%           | 84.44%           | 83.50%           |
| Recall          | 100.00%          | 17.03%           | 31.55%           | 44.86%           | 57.08%           | 63.86%           | 69.34%           | 71.81%           | 74.24%           | 76.92%           | 78.75%           |
| Reproductive    | mAP 69.74%       | 64.81%           | 73.03%           | 77.30%           | 77.05%           | 76.16%           | 75.55%           | 74.86%           | 74.20%           | 73.60%           | 73.02%           |
| Recall          | 100.00%          | 22.44%           | 43.06%           | 59.33%           | 68.32%           | 74.26%           | 78.36%           | 81.68%           | 84.44%           | 86.70%           | 88.74%           |
cannot be explained, such evaluation is not feasible in practice.\textsuperscript{[23]} Our system, however, enables the users to understand the basis for diagnostic recommendations and provide doctors with comprehensive and accurate diagnostic support, even when the doctors are not equipped with AI knowledge. Specifically, the knowledge graph can visually display the relationship between the lab data and the diagnostic data; the MiXG model can present each feature included in the model and the importance scores of these features. The combination of the two technologies can greatly improve the transparency and interpretability of the AI system. Therefore, this comprehensive analysis method can help to achieve full coverage of high-level laboratory-based diagnosis.

As we all know, advanced imaging and pathological tests are less available in resource-constrained settings, while laboratory tests are the most commonly used diagnostic method.\textsuperscript{[24]} We envision that AI LEON could be used as an assistive tool by primary healthcare (PHC) doctors. Pertaining to the high top-$n$ accuracy of the model, AI LEON could be used to narrow down the potential disease diagnosis using the top-$n$ predicted diagnoses, followed by more informed ancillary tests to obtain the final diagnoses. By doing this, the AI LEON-assisted clinical workflow could be optimized, and doctors’ misdiagnosis and missed diagnosis rate may also be reduced. In addition, in developing regions with a limited number of experienced doctors, our

---

**Figure 5.** Clinical application scenarios of AI LEON. a) Gold standard-based diagnosis using knowledge-driven models based on the guidelines and expert consensus. Using COVID-19 as an example, the infection of COVID-19 could be confirmed by COVID-19 diagnostic formula, evaluating COVID-19 RNA testing report with at least two positive genes of ORF1ab, N and E. b) Specific diagnosis using etiology-related knowledge-driven models. Using renal anemia as an example, the definitive diagnosis was finally made by integrating the anemia diagnostic formula and CKD formula. c) Differential diagnosis by combining knowledge-driven models with data-driven models. Using CA19-9 as an example, the knowledge-driven model first suggests that the increase in CA19-9 may be related to some diseases, such as pancreatic cancer, pancreatitis, ovarian cysts, etc. Then, the corresponding data-driven models for each disease were triggered and further the positivity and relevance of these diseases were calculated. At the same time, the importance scores of features included in each model were presented. Abbreviations: AI LEON, Artificial Intelligence-based Lab Est to diagnosis; CKD, chronic kidney disease; RT, routine test.
Figure 6. Comparison of the diagnostic performance between AI LEON and human physicians. 
a) The diagnostic performance of AI LEON and three GP groups in the blank-filling study for overall patients. 
b–k) The diagnostic performance of AI LEON and three GP groups in the blank-filling study for ten organ-based subgroups of patients. 
l) The diagnostic performance of AI LEON and three GP groups in the true or false study for overall patients. 
m–v) The diagnostic performance of AI LEON and three GP groups in the true or false study for ten organ-based subgroups of patients. 
Abbreviations: AI LEON, Artificial Intelligence-based Lab Test to diagnosis; GPs, General Practitioners; mAP, mean average precision.
Table 5. Comparison of diagnostic performance between AI LEON and three groups of GPs in the overall patients.

|                  | Accuracy |          | mAP     |          | Recall  |
|------------------|----------|----------|---------|----------|---------|
|                  | Mean     | SD       | Mean    | SD       | Mean    |
| Blank filling    | 0.791    | —        | 0.862   | —        | 0.919   |
| GP group 1       | 0.637    | 0.066827 | 0.183   | 0.070427 | 0.165   | 0.068562|
| GP group 2       | 0.487    | 0.221956 | 0.183   | 0.063219 | 0.168   | 0.063185|
| GP group 3       | 0.499    | 0.144001 | 0.176   | 0.029813 | 0.162   | 0.026711|
| True—false       | 0.759    | —        | 0.832   | —        | 0.909   |
| GP group 1       | 0.757    | 0.011589 | 0.455   | 0.116385 | 0.451   | 0.116225|
| GP group 2       | 0.7      | 0.120804 | 0.456   | 0.072535 | 0.447   | 0.078883|
| GP group 3       | 0.754    | 0.052257 | 0.463   | 0.090953 | 0.459   | 0.098241|

system could be deployed at the lowest cost to provide direct medical consultation to patients (such as mobile APPs). Insufficient medical resources and the continuous increase in chronic diseases are challenging our medical systems.[25] Our real-time AI diagnostic system could provide patients with preliminary diagnosis results and perform appropriate hierarchical diagnosis and treatment, thereby alleviating the shortage of medical resources. It has been well documented that the benefits from the patient’s laboratory objective data involve screening the population with asymptomatic complex diseases, risk stratification of diseases, specific diagnosis of the patient’s condition, etc.[19] Therefore, we expect that our laboratory-based diagnostic system can help improve the early diagnosis of complex and refractory diseases, increasing the survival rates and decreasing the treatment expense.

Despite these encouraging results, our AI system still has some limitations. First, the knowledge graph base needs to be iterated and updated continuously to ensure that our system always contains the latest medical laboratory knowledge. Secondly, as the accuracy of our system is heavily dependent on the quality of the input data, all the laboratory data should be carefully preprocessed. Also, if our system is to be generalized to an external test cohort, the harmonization of lab data from different healthcare systems is the key premise. By standardization and harmonization of laboratory methods and results,[16,19] these limitations could probably be lifted.

We have established a knowledge-and-data-driven intelligent laboratory system, AI LEON, with good robustness and interpretability, achieving a full high-level laboratory-based diagnosis coverage. Our study provides a proof of concept for implementing a laboratory-based intelligent system to imitate and even augment the clinical diagnosis capabilities of human physicians. AI LEON can ultimately speed up the accurate diagnosis and timely referral of complex and refractory diseases, thereby facilitating earlier treatment and improving clinical outcomes. Therefore, our tool is expected to be applicable to clinical diagnosis and particularly useful for areas with limited medical resources.

4. Experimental Section

Study Approval: The study protocol was reviewed and approved by the Institutional Ethics Committee at Shanghai Changhai Hospital.

Study Design and Participants: An ambispective diagnostic study was conducted in Shanghai Changhai Hospital. Eligible criteria of the patients between Jan 1, 2010, and Oct 1, 2020, were listed as follows: 1) having diagnostic results from the ICD versions 9 or 10 and 2) having laboratory test data within 21 days prior to diagnosis. Exclusion criteria included 1) follow-up diagnosis beyond the initial diagnosis to minimize treatment bias and 2) diagnosis without correlation with laboratory medicine because many diseases are determined by pathology or imaging (such as skin cancer, fractures). Eligible participants between Jan 1, 2010, and Dec 31, 2019, were included in the retrospective cohort, and those diagnosed between Jan 1, 2020, and Oct 1, 2020, were included in the prospective test cohort. The retrospective cohort was randomly divided into a training set and validation set (8:2) to develop and validate the AI LEON system, respectively. Then, the system was prospectively assessed in the test cohort.

Data Source: To improve clinical scientific research, Shanghai Changhai Hospital developed an integrated data acquisition system and built a large-scale data integration platform including Clinical Data Repository (CDR) and Research Data Repository (RDR),[26] which enforced data mapping to controlled vocabularies such as Systematized Nomenclature of Medicine Clinical Term (SNOMED CT), the ICD versions 9 and 10, Logical Observation Identifiers Names and Codes (LOINC), Current Procedural Terminology (CPT), Healthcare Common Procedure Coding System (HCPCS), etc. We also developed a data extraction package specific to the data integration platform of Shanghai Changhai Hospital, extracted data comprising anonymized demographics, laboratory test results, diagnoses information, medical records, pathology, and radiology reports of eligible participants.

Data Processing: 1) Data Standardization: There were three data types in a test result, quantitative, qualitative, and character. We thus ensured that all data with different types were unified and standardized. 2) Data Normalization: Our normalization process included standardizing continuous test results and setting a finite boundary for results that might go to infinity, solving the infinite left or right boundary of the reference range. From the perspective of mathematical modeling, our normalization method adopted a nonequivalent system change. Moreover, the normalized result had a symmetrical structure that was convenient for modeling and analysis. 3) Time Window: As both the independent and dependent variables have time-series characteristics, it was necessary to combine the independent and dependent variables to form a sample. We chose
a time window of $T = 21$ days. Thus, the independent variables were collected within 21 days before the time point of the dependent variable. 4) Missing Value: Missing values in the data corresponded to cases where the doctor did not recommend the patient to have these tests, which may be within the normal range according to the doctor’s initial inquiry. Therefore, we filled in the missing values as “negative.”

The knowledge-driven Models: 1) Establishment of Knowledge-driven Models: Senior specialists manually annotated a total of the 3156 diagnostic flowchart based on 24693 knowledge items. Then, the knowledge-driven models were established based on the knowledge graph analysis framework (https://www.ontotext.com/knowledgehub/fundamentals/how-to-building-knowledge-graphs-in-10-steps/).

### Table 6. Comparison of diagnostic performance between AI LEON and three groups of GPs in organ-based subgroups of patients.

| System                | Blank-filling study | True–false study |
|-----------------------|---------------------|------------------|
|                       | AI LEON  | GP group 1  | GP group 2 | GP group 3 | AI LEON  | GP group 1 | GP group 2 | GP group 3 |
| Infection             |          |             |            |            |          |             |            |            |
| Accuracy              | 0.792    | 0.685       | 0.557      | 0.505      | 0.758    | 0.78       | 0.762      | 0.724      |
| Recall                | 0.918    | 0.287       | 0.234      | 0.24       | 0.911    | 0.588      | 0.6         | 0.579      |
| mAP                   | 0.862    | 0.309       | 0.256      | 0.254      | 0.833    | 0.592      | 0.606      | 0.579      |
| Respiratory system    |          |             |            |            |          |             |            |            |
| Accuracy              | 0.784    | 0.5         | 0.6        | 0.375      | 0.758    | 0.667      | 0.615      | 0.524      |
| Recall                | 0.921    | 0.2         | 0.444      | 0.164      | 0.925    | 0.513      | 0.429      | 0.355      |
| mAP                   | 0.866    | 0.219       | 0.472      | 0.185      | 0.836    | 0.467      | 0.456      | 0.374      |
| Hematological system  |          |             |            |            |          |             |            |            |
| Accuracy              | 0.795    | 0.639       | 0.276      | 0.317      | 0.755    | 0.747      | 0.614      | 0.699      |
| Recall                | 0.922    | 0.194       | 0.142      | 0.106      | 0.916    | 0.487      | 0.477      | 0.514      |
| mAP                   | 0.878    | 0.204       | 0.164      | 0.12       | 0.832    | 0.485      | 0.472      | 0.512      |
| Gastrointestinal system|        |             |            |            |          |             |            |            |
| Accuracy              | 0.797    | 0.678       | 0.216      | 0.52       | 0.759    | 0.764      | 0.558      | 0.716      |
| Recall                | 0.92     | 0.27        | 0.235      | 0.207      | 0.906    | 0.507      | 0.548      | 0.535      |
| mAP                   | 0.873    | 0.286       | 0.316      | 0.223      | 0.833    | 0.513      | 0.559      | 0.532      |
| Urinary system        |          |             |            |            |          |             |            |            |
| Accuracy              | 0.785    | 0.684       | 0.424      | 0.461      | 0.747    | 0.757      | 0.716      | 0.712      |
| Recall                | 0.903    | 0.272       | 0.209      | 0.178      | 0.904    | 0.544      | 0.548      | 0.534      |
| mAP                   | 0.846    | 0.292       | 0.215      | 0.182      | 0.822    | 0.543      | 0.538      | 0.525      |
| Endocrine system      |          |             |            |            |          |             |            |            |
| Accuracy              | 0.793    | 0.695       | 0.216      | 0.566      | 0.759    | 0.767      | 0.591      | 0.729      |
| Recall                | 0.921    | 0.275       | 0.233      | 0.214      | 0.908    | 0.509      | 0.574      | 0.55       |
| mAP                   | 0.867    | 0.293       | 0.311      | 0.226      | 0.831    | 0.51       | 0.588      | 0.55       |
| Circulatory system    |          |             |            |            |          |             |            |            |
| Accuracy              | 0.796    | 0.664       | 0.591      | 0.516      | 0.756    | 0.758      | 0.731      | 0.709      |
| Recall                | 0.921    | 0.259       | 0.319      | 0.232      | 0.913    | 0.535      | 0.601      | 0.571      |
| mAP                   | 0.862    | 0.278       | 0.348      | 0.243      | 0.829    | 0.531      | 0.605      | 0.571      |
| Immune system         |          |             |            |            |          |             |            |            |
| Accuracy              | 0.797    | 0.6         | 0.358      | 0.494      | 0.757    | 0.739      | 0.631      | 0.678      |
| Recall                | 0.924    | 0.212       | 0.199      | 0.265      | 0.935    | 0.567      | 0.463      | 0.526      |
| mAP                   | 0.851    | 0.255       | 0.045      | 0.266      | 0.880    | 0.57       | 0.415      | 0.486      |
| Reproductive system   |          |             |            |            |          |             |            |            |
| Accuracy              | 0.857    | 0.625       | 0.364      | 0.4        | 0.818    | 0.833      | 0.727      | 0.64       |
| Recall                | 0.933    | 0.109       | 0.154      | 0.143      | 0.9      | 0.5        | 0.381      | 0.4        |
| mAP                   | 0.899    | 0.122       | 0.114      | 0.139      | 0.86     | 0.467      | 0.37       | 0.359      |
| Nervous system        |          |             |            |            |          |             |            |            |
| Accuracy              | 0.813    | 0         | 0         | 0.125      | 0.667    | 0          | 0          | 0          |
| Recall                | 0.867    | 0         | 0         | 0.024      | 1        | 0          | 0          | 0          |
| mAP                   | 0.76     | 0         | 0         | 0.033      | 0.976    | 0          | 0          | 0          |
Our knowledge-driven models partitioned the knowledge corresponding to test results into two categories, defined knowledge and ambiguous knowledge. Defined knowledge shows very strong confidence in supporting a diagnosis based on a combination of some test results. As an example of defined knowledge, the infection of COVID-19 could be confirmed with very high confidence by its RNA testing report with at least two positive genes of ORF1ab, N and E according to the protocol of the COVID-19 test kit. Ambiguous knowledge shows a possibility that some suspicious diseases might be the result of certain abnormal test results. As an example of ambiguous knowledge, a high test CA19-9 value may be the evidence of pancreatic cancer, pancreatitis, ovarian cyst, etc. However, we need further evidence to finalize the diagnosis.

2) Model Expression Paradigm: The data included test results ($S_1$), diagnostic results ($S_2$), and basic information results ($S_3$). Test results included test specimen, name of the test item, result, unit, reference range, normalized result, test time. Diagnosis results included diagnosis name, diagnosis ICD10 code, and diagnosis time. Basic information result included gender, birth year, and race.

The model inputs were obtained from the above raw data, which included test parameters ($P_1$), diagnostic parameters ($P_2$), and basic information parameters ($P_3$). Test parameters included test specimen and name of the test item; diagnostic parameters only contained the name of the diagnosis; and basic information parameters included gender, birth date, and race.

In the model, $S$ represented the patient's actual test results as the arguments passed to a function; $P$ represented the parameters needed for the model, similar to the parameters in a function. For example, for a patient with a typical CA19-9 lab result, this lab result was taken from $S_1$; but in a knowledge model, an elevated CA199 will alert pancreatic cancer; then elevated CA199 was $P_1$, which represented a parameter needed by the model. The overall knowledge inference process continuously compares the actual arguments in $S_1$ with the morphological parameters in $P_1$ to find the conditions that can satisfy the knowledge model and complete the knowledge inference.

The model took in an input and gave an output that shows the confidence level of a diagnosis. High confidence in the output indicates the gold standard. Medium-level confidence indicates that the knowledge inference results can be directly displayed to the physician. Low-level confidence means that the result needs the data-driven models for a secondary judgment.

3) The Use of the Model: The use of the model is as follows.

Step 1: For a patient’s clinical dataset $(List < \hat{s}_1 >, List < \hat{s}_2 >, \hat{s}_3 >)$, we generated the corresponding input $(List < \hat{p}_1 >, List < \hat{p}_2 >)$.

- Initially set $M_0 = \Phi$ and $L = \Phi$.
- Step 2: Among all knowledge models, we found the set of knowledge models that met the requirements.

\[ M_1 = \{ \text{model}_i | List < \hat{p}_1 > \subseteq List < \hat{p}_1 > \text{and List} < \hat{p}_2 > \subseteq List < \hat{p}_2 > \text{and model}_i \not\in M_0 \} \]  

(1)

If $M_1 = \Phi$, go to Step 5.
Otherwise, go to Step 3.

Step 3: $M_1$ included three types of models with confidence level: high, medium, and low, which were noted as $M_{1H}, M_{1M},$ and $M_{1L}$, respectively.

The models in the knowledge model collections $M_{1H}, M_{1M},$ and $M_{1L}$ were operated separately, and the inference results of the same level models were merged. The merged outputs were recorded as $List < \text{output}_1 >, List < \text{output}_2 >,$ and $List < \text{output}_3 >$.

Step 4: We combined the results in $List < \text{output}_i >$, into $L$.

The results of $List < \text{output}_1 >$, and $List < \text{output}_2 >$, and $List < \text{output}_3 >$ were merged with $(List < \hat{s}_1 >, List < \hat{s}_2 >, \hat{s}_3 >)$ and denoted as $(List < \hat{s}_1 >, List < \hat{s}_2 >, \hat{s}_3 >)$.

Then, a corresponding parameter list $(List < \hat{p}_1 >, List < \hat{p}_2 >)$ was generated.

Step 5: The result of $(List < \hat{s}_1 >, List < \hat{s}_2 >, \hat{s}_3 >)$ was the final model inference result, and the $List < \hat{s}_2 >$ was the diagnosis result with high and medium-inference levels, while the diagnosis information contained in the $L$ set was the diagnosis result with low-inference basis. The model inference process used the original knowledge in the recorded $M_1$ as the inference basis.

A graphical representation of data flows within the model is shown in Figure 7.

The Data-Driven Models: 1) Establishing the Extreme Gradient Boosting (XGBoost) Model: For each disease, we used the classical XGBoost method to construct a binary classification model. As an ensemble method using boosted decision trees, XGBoost is famous for both speed and performance.

The objective function $\mathcal{L}(t)$ at step $t$ was given as

\[ \mathcal{L}(t) = \sum_{i=1}^{N} \left( f_i(y_i) - f_i(x_i) \right)^2 + \Omega(f_i) \]

(2)

where the first term is the loss function that sums the loss of actual value $y_i$ and its predicted value $\hat{y}_i^{(t-1)} + f_i(x_i)$, and the second term is a regularization penalty. At step $t$, the model aims to find the best function $f_i$ to be added to the current model. Its advantages include both performance and speed. Also, like other boosting methods, even though one could not directly obtain the contribution of each feature to the final prediction, an importance score could be obtained for each feature. This importance score provided interpretability to the automated diagnosis, assisting the doctors in making decisions.

2) Establishing the Genetic algorithm (GA) Model: We used the GA to choose the near-optimal combination of thresholds of all diseases to maximize the mAP metric. GA is a search heuristic inspired by Darwin’s natural evolution theory, where the genes of the fittest individuals are inherited by the offsprings and genes that are less competitive are discarded in the evolution. In our problem, we represented genes by the threshold values of the units, and the fitness function was the mAP that measured the goodness of a ranking algorithm. It should be noted that the diagnosis is more of a ranking process than a classifying one. Compared with multi-label classifications, the ranking process takes into account both classification accuracy and diagnosis relevancy. An expert doctor usually writes his or her diagnoses in decreasing order of relevance and severity. We used the mAP metric to ensure that the most relevant and severe diagnostic result was listed atop.

The genes of our algorithm were the threshold values that determined the binary diagnostic classes from the outputs of the ML algorithms. A chromosome is simply the collection of all threshold values for all outputs. We chose an initial population size of 100 and ran a total of 10 iterations. For each pair of parents to be mated, a crossover point was chosen randomly from within the genes. The new offspring was created by exchanging the genes of the parents divided by the random crossover point. We then added the new offspring to the population and removed the ones with the lowest fitness values.

We selected a random crossover from the genes for each pair of parents. New offsprings were created by swapping the corresponding genes of the parents according to the crossover. Then, we replaced the offspring having the lowest fitness value with the new offspring.

3) Modeling with MIXG: Our data-driven model is a ranking system that mimics human doctors. Therefore, the proposed MIXG algorithm first calculates the probabilities of different diseases from the XGBoost classifiers and then ranks them according to the results from GA models. Here, the standard evaluation metrics for classification, such as out-of-sample misclassification error, the area under the receiver operating characteristic curve (ROC–AUC), and the sensitivity and specificity, could not be used.
Data Normalization

Knowledge-Driven Model Screening: Match S1, S2, S3 with P1, P2, P3 in the knowledge model to find the proper knowledge models

Inference with Screened Knowledge-Driven Model

Output Result: Return Data-Driven model: Return Probabilities for each Disease

Rectify the output probability by the genetic algorithm

Return a list ranked by the rectified probability

Output Result

as the appropriate evaluation metrics. Instead, our system used a mAP function as the evaluation metric for the ranking process. The mAP formulation was formally written as

\[ \text{mAP} = \frac{1}{M} \sum_{m=1}^{M} \text{AP} \]  \hspace{1cm} (3)

where \( \text{AP} \) is the average precision defined as

\[ \text{AP} = \int_{0}^{1} p(r)dr \]  \hspace{1cm} (4)

In the earlier formula, \( p(r) \) is the precision \( p \) as a function of recall \( r \). Recall is defined as the percentage of positive diagnoses that were retrieved. Mathematically

\[ \text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \]  \hspace{1cm} (5)

where TP is the total number of true positives and FN is the number of false negatives. The sum of TP and FN constitutes all positive diagnoses in the dataset. Recall increased as we went down the ranking list. The precision did not necessarily decrease but depended on the ranking algorithm’s accuracy. Thus, the function might not be monotonically decreasing and might have a zigzag pattern. Therefore, the integral was in practice replaced with a finite sum of interpolated terms \( p^*(r) \), where \( p^*(r) = \max_{0 \leq k \leq r} p(r) \) that replaced the precision value at a recall value \( r \) with the maximum precision to the right of that recall value. Thus, the average precision was calculated as

\[ \text{mAP} = \frac{1}{M} \sum_{k=1}^{M} p^*(r_k) \Delta r(k) \]  \hspace{1cm} (6)

where \( \Delta r(k) \) is the change in recall from ranking \( k-1 \) to \( k \). The formula is a general interpolated average precision function for a list of precision values and recall values from a given ranked list.

Integration of Knowledge-Driven and Data-Driven Models: We combined the knowledge-driven and data-driven models to build AI LEON, a test result diagnosis engine that provided evidence-based medical evidence and used past experiences learnt from the data-driven models. The basic working mechanism of LEON was that it applied the knowledge-driven model and obtained the results based on high, medium, or low confidence from learnt knowledge. The knowledge was important in both data filtration for better machine understandability and diagnosis for improved results’ interpretability. At the same time, LEON initiated a data-driven model using the MiXG algorithm for diagnoses, obtained probability scores for each disease, and ranked the diagnoses in the order of probability from the highest to the lowest to form an ordered list of diagnostic risks.

Competition between AI LEON and GPs in Reading Lab Reports: We also compared the diagnostic performance of AI LEON with that of 15 junior GPs in reading lab reports of 100 cases from the test set. GPs are health service providers with a high degree of comprehensiveness and perform general practice covering a variety of medical problems in patients of all ages. Fifteen junior GPs were randomly divided into three groups. 100 laboratory test reports (lab reports) were randomly selected from the test set and then assigned to each group of GPs. All the GPs read the lab reports and recorded the list of diagnoses. The reading consisted of two parts, a blank-filling study and a true–false one. In the blank-filling study, readers were asked to write down the list of diagnoses. In the true–false study, readers were asked to determine whether the machine-recommended diagnoses were correct. Laboratory test reviews conducted by physicians as well as AI LEON were all done without clinical information. Then, ten specialists were invited to provide diagnostic results of the 100 cases as the gold standards for human—machine tests. The accuracy of each diagnosis was recorded according to the gold standards. mAP was utilized to evaluate the accuracy of multiple diagnoses, and recall value was used to assess the rate of missed diagnoses.

Data Availability: The clinical data used for the training, validation, and test sets were collected from the data integration platform of Shanghai Changhai Hospital with strict access control in a nonidentifiable format. The use of the data was locally licensed and ethically approved. It is not publicly available, and there are restrictions on its use. However, all researchers can contact the corresponding author and request access to the relevant data.

Code Availability: We used several open-source R libraries (e.g., xgboost) to conduct our experiments. The analysis was performed.
with code written in Python 3.5. Source code for the data-driven MixG model is available under an open-source license at https://github.com/hitaless-tech/MixG. We described the details of experiments in the Experimental Section to allow for independent replication. Further inquiries regarding the specific nature of the models can be made by relevant parties to the corresponding author.

Acknowledgements

B.W., J.J., X.H., and C.H. contributed equally to this work. The authors thank Shanghai Science and Technology Committee Program (20511101200) for financial support. Thanks to the clinicians from different Departments of Shanghai Hospital: Prof. Xianbao Zhan, Dr. Yingyi Zhang, Dr. Jie Li, Dr. MeiHong Wu (Oncology); Prof. Xiaoying Bi, Dr. Chunhua Feng, Dr. Shu Zhou, Dr. Zhenyiu Liu, Dr. Youbei Liu (Neurology); Prof Zhiyong Guo, Dr. Juan Li, Dr. Qi Bian, Dr. Jing Sun (Nephrology); Dr. Jie Gao, Dr. Dongbao Zhao, Dr. Wei Wan, Dr. Lianmei Ji, Dr. Ju Zhang (Rheumatology); Dr. Xin Zhang, Dr. Rui Zheng (General Surgery); Prof. Yan Shang, Dr. Jiayi Zhao, Dr. Junjie Wu, Dr. Yang Jiao (General Medicine); Dr. Lei Li, Dr. Xiangyu Kong (Gastroenterology); Prof. Maojin Xu (Endocrinology); Prof. Fubo Wang (Urology); and Dr. Xiaoliang Cong (Cardiology, Shanghai Changzheng Hospital). The authors also want to thank Dr. Zhang Jianfeng (Nantong Tumor Hospital); Dr. Yunhan Ji (Orthopedic Surgery Tongren Hospital); Prof. Cheng Wu (Department of Health Statistics); Prof. Jun Zheng and Prof. Wenxin Hu (School of Data Science & Engineering, East China Normal University); and Mr. Chao Xiong and Mr. Nan Zhang (Hitale, Shanghai) for their professional guidance and support to the establishment of the AI LEON. Moreover, thanks to all the general practitioners from the standardized training organization of general medicine who participated actively in the Examination Report Interpretation Challenge with AI LEON.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

B.W., J.J., X.H., C.H., F.W., C.M., and S.L. were involved in the study design. Q.Q., Y.J., and Z.W. collected and processed the data. L.J., B.G., I.W., and X.Z. contributed to the analysis methods. B.W., J.J., H.X., F.W., C.M., and S.L. analyzed and interpreted the data. B.W., J.J., C.H., F.W., C.M., and S.L. contributed to the writing of the report. All the authors approved the corresponding author upon reasonable request.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

artificial intelligence, diagnosis, knowledge and data driven, laboratory medicine, machine learning

[1] A. Rajkomar, J. Dean, I. Kohane, N. Engl. J. Med. 2019, 380, 1347.
[2] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, S. Thrun, Nature 2017, 542, 115.
[3] D. Kermany, M. Goldbaum, W. Cai, C. Valientim, H. Liang, S. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. Prasadha, J. Pei, M. Ting, J. Zhu, C. Li, S. Hewett, I. Ziyar, A. Shi, R. Zhang, L. Zhang, R. Hou, W. Shi, W. Chuan, V. Yuan, V. Huan, W. Chen, E. Zhang, C. Zhang, O. Li, et al., Cell 2018, 172, 1122.
[4] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Muller, WIREs Data Min. Knowl. Discov. 2019, 9, 1312.
[5] M. Mehta, I. Katz, A. Jha, N. Engl. J. Med. 2020, 382, 791.
[6] A. Beam, A. Manrai, M. Ghassemi, Nature 2020, 577, 2402.
[7] A. Holzinger, C. Reinhold, D. Martin, R. Forghani, Neuroimaging Clin. N. Am. 2020, 30, 401.
[8] B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, M. Ahmadian, M. van Engen-Verheul, F. Bakhshi-Raiez, N. Peek, J. R. Genzen, C. D. Burnham, R. A. Felder, C. D. Hawker, G. Lippi, A. Holzinger, A. Bobicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, Nat. Med. 2019, 25, 24.
[9] P. Savadjiev, C. Reinhold, D. Martin, R. Forghani, Neuroimaging Clin. N. Am. 2020, 30, 401.
[10] V. Gulshan, L. Peng, M. Coram, M. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. Nelson, J. Mega, D. Webster, JAMA 2016, 316, 2402.
[11] R. Gargaya, T. Leng, Ophthalmology 2017, 124, 962.
[12] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, W. Shi, R. Hou, W. Shi, W. Chuan, V. Yuan, V. Huan, W. Chen, E. Zhang, C. Zhang, O. Li, et al., Cell 2018, 172, 1122.
[22] E. H. Shortliffe, M. J. Sepúlveda, JAMA 2018, 320, 2199.

[23] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, K. Zhang, Nat. Med. 2019, 25, 30.

[24] D. A. Orringer, B. Pandian, Y. S. Niknafs, T. C. Hollon, J. Boyle, S. Lewis, M. Garrard, S. L. Hervey-Jumper, H. J. L. Garton, C. O. Maher, J. A. Heth, O. Sagher, D. A. Wilkinson, M. Snuderl, S. Venneti, S. H. Ramkissoon, K. A. McFadden, A. Fisher-Hubbard, A. P. Lieberman, T. D. Johnson, X. S. Xie, J. K. Trautman, C. W. Freudiger, S. Camelo-Piragua, Nat. Biomed. Eng. 2017, 1, 0027.

[25] S. Jan, T. L. Laba, B. M. Essue, A. Cheorghe, J. Muhunthan, M. Engelgau, A. Mahal, U. Griffiths, D. McIntyre, Q. Meng, R. Nugent, R. Atun, Lancet 2018, 391, 2047.

[26] B. Wang, W. Wei, Z. Shao, Q. Qin, Z. Wang, Y. Jia, J. Guo, Y. Pang, L. Jiang, G. Jin, C. Mao, S. Liu, Adv. Intell. Syst. 2021, 3, 2100033.