Article

A Preliminary Evaluation of "GenDAI", an AI-Assisted Laboratory Diagnostics Solution for Genomic Applications [†]

Thomas Krause [1,*] , **Elena Jolkver** [1] , **Sebastian Bruchhaus** [1] , **Paul Mc Kevitt** [2] , **Michael Kramer** [3] **and Matthias Hemmje** [2]

1 Faculty of Mathematics and Computer Science, University of Hagen, 58097 Hagen, Germany; elena.jolkver@studium.fernuni-hagen.de (E.J.); sebastian.bruchhaus@fernuni-hagen.de (S.B.)

2 Research Institute for Telecommunication and Cooperation (FTK), 44149 Dortmund, Germany; pmckevitt@ftk.de (P.M.K.); mhemmje@ftk.de (M.H.)

3 ImmBioMed Business Consultants GmbH & Co. KG, 64319 Pfungstadt, Germany; m.kramer@immbiomed.de

* Correspondence: thomas.krause@fernuni-hagen.de

† This paper is an extended version of our paper published in the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2021), Machine Learning and Artificial Intelligence in Bioinformatics and Medical Informatics (MABM 2021) Workshop, 9–12 December 2021.

Abstract: Genomic data enable the development of new biomarkers in diagnostic laboratories. Examples include data from gene expression analyses or metagenomics. Artificial intelligence can help to analyze these data. However, diagnostic laboratories face various technical and regulatory challenges to harness these data. Existing software for genomic data is usually designed for research and does not meet the requirements for use as a diagnostic tool. To address these challenges, we recently proposed a conceptual architecture called "GenDAI". An initial evaluation of "GenDAI" was conducted in collaboration with a small laboratory in the form of a preliminary study. The results of this pre-study highlight the requirement for and feasibility of the approach. The pre-study also yields detailed technical and regulatory requirements, use cases from laboratory practice, and a prototype called "PlateFlow" for exploring user interface concepts.

Keywords: laboratory diagnostics; requirements engineering; genomics; gene expression; metagenomics; medical diagnostics

1. Introduction

The field of biology that focuses on the genetic material of organisms is known as genomics. The goal of genomics is to understand the function, mechanisms, and regulation of genes and other genomic elements. This includes, for example, understanding the complex relationship between the genome, the expression of individual genes, environmental factors, and the resulting physiological and pathophysiological states of a cell or organism as a whole.

Once these relationships are discovered, they can be used in laboratory diagnostics to identify pathophysiological conditions and potentially improve patient care. Two examples are gene expression analysis and metagenomic analysis. The former is used to measure the activity of certain genes, while the latter is used to analyze the genomes of a patient's microbiota (e.g., the gut microbiota). Medical laboratories must constantly evolve and adapt to bring new insights into clinical practice. This includes generating, transforming, combining, and evaluating data to deliver individualized diagnostic results.

Genomic applications often generate large quantities of data [1], which presents processing and analysis challenges. For example, human genome sequencing regularly generates hundreds of gigabytes of raw data for a single sample. It is foreseeable that the total quantity of data will continue to increase as new technologies are developed and existing technologies are used more extensively. The type of data generated in these genomic

applications is also quite heterogeneous and they sometimes need to be combined with other available data in the context of personalized medicine. This combination of volume (of data), velocity (of increase in data), and variety, identifies genomic applications as a Big Data problem [2]. Big Data application requirements often exceed the limits of individual machines and thus depend on architectures that are scalable across machine boundaries.

Artificial intelligence (AI) can identify patterns in such large data sets. In particular, "deep learning" [3] has proven to be a powerful technique for detecting even complex relationships in data. It has found application in a number of complex problems, such as phenotype prediction and regulatory genomics [4]. Due to its increased computational power, this technique can process large quantities of data. At the same time, deep learning often requires less data preprocessing than other approaches. For AI to support data analysis, several other challenges must be addressed, such as model selection, feature engineering, model explainability, and reproducibility, which are exacerbated by high dimensionality and a relatively small number of samples, often referred to as the "curse of dimensionality" [5]. In laboratory diagnostics, too, the number of samples available for AI methods is usually limited, since obtaining a larger number of samples is often complex and expensive. AI is also applied beyond analysis [5,6]. Examples include its use for dimensionality reduction in the visualization of high-dimensional data or clustering of similar sequences [7].

Regardless of the application area, the explainability of AI models is a further challenge. Powerful methods such as deep learning often represent a "black box", where it is unclear according to the criteria through which the model arrives at a certain decision. This is not only a purely technical challenge, but also a regulatory challenge in which legislators and regulators must create clear criteria according to which the use of AI in laboratory diagnostics is permitted. Initial proposals in this regard were published, for example, in 2020 by the Joint Research Center (JRC) of the European Commission [8]. In this report, transparency, reliability, and data protection are named as core criteria against which AI models must be measured.

In the context of laboratory diagnostics, genomic applications and AI face additional regulatory and technical challenges. Applicable standards and regulations such as ISO standards (including ISO 13485 [9], ISO 15189 [10], and IEC 62304 [11]) and the European Union's In Vitro Diagnostics Regulation (IVDR) [12] require that instruments used in diagnostics, as well as software, be certified and meet numerous criteria.

"Health institutions", defined in the IVDR as "…an organization the primary purpose of which is the care or treatment of patients or the promotion of public health" [12], have the privilege of using so-called "Laboratory-Developed Tests (LDTs)" [13]. For these tests, the laboratory takes full responsibility for validation and IVDR compliance. These tests may require the application of laboratory-developed software to convert raw data into reportable results. Taking responsibility for conformity with IVDR requirements means that the health institution has to establish and document that the LDT complies with the essential safety and performance requirements as specified in Annex I of the IVDR. Moreover, the institution has to establish and operate a Quality Management System (QMS) [10], including a Risk Management System (RMS) [13], Post-Market Surveillance (PMS), and Post-Marketing Performance Follow-Up (PMPF) for the respective LDTs. PMS and PMPF are intended to make sure that new scientific findings or technical developments are recognized, taken into account, and—when appropriate—implemented even after the tests have been introduced. This is to ensure that the performance of diagnostic tests always reflects the current state of the art. For software used in conjunction with LDTs, the same requirements apply as this software was not approved by the manufacturer for diagnostic purposes. It may thus be termed "RUO software" (Research Use Only).

However, because RUO applications have not been optimized for laboratory diagnostics, they may be difficult to integrate into the laboratory workflow, reducing efficiency and thus increasing costs. An example of this is software that is based on the concept of projects or individual experiments rather than automating repetitive tasks. Another example is data

transfer between the RUO software and other parts of the solution, such as the compliance systems mentioned earlier.

Whether the required software is being developed from scratch or existing software is being repurposed for the task, careful analysis of the requirements for such software must be undertaken to ensure that a solution complies with all applicable regulations, supports laboratory use cases, integrates with other relevant systems, and does all this in an efficient manner, automating processes where possible to reduce the possibility of errors and overall costs. Due to these numerous aspects, involving different areas such as biology, informatics, and regulations, requirements engineering for laboratory diagnostic software is challenging.

In summary, challenges arise from the constant advancement of science and technology, the quantity and type of data, the use of machine learning to process these data, regulations governing the laboratory process, and the identification and analysis of requirements for laboratory diagnostic software. Combining genomic applications with AI and applying them in the context of laboratory diagnostics has potential for improved diagnostics, but only if the above challenges can be overcome (Figure 1).
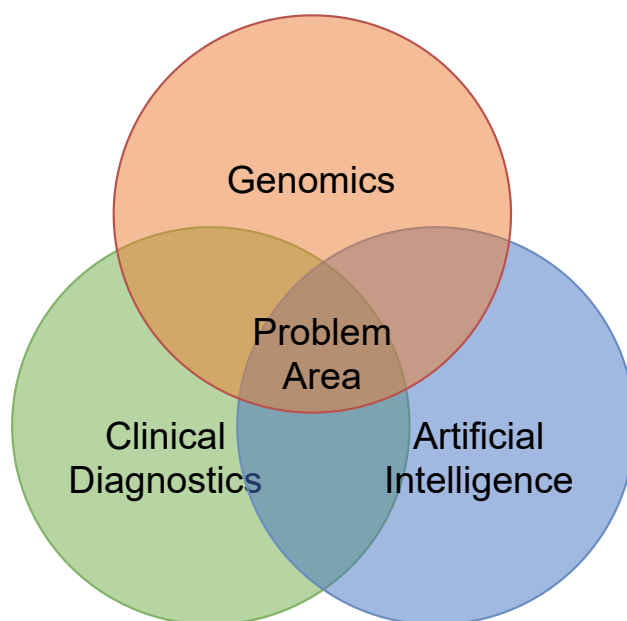

Figure 1. Problem area adressed in this paper, © 2021 IEEE. Reprinted, with permission, from Krause et al. [14].

Towards this goal, we recently introduced [14] a conceptual model called "GenDAI" (GENomic applications for laboratory Diagnostics supported by Artificial Intelligence). This extended paper discusses the rationale behind GenDAI in more detail and provides an initial evaluation of the model with the help of a recently conducted pre-study. The remainder of the paper discusses existing conceptual models that were developed before "GenDAI". "GenDAI" is then introduced as a new conceptual model that was developed to comprehensively cover the outlined challenges and the use cases in laboratory diagnostics. Finally, we discuss the pre-study, exploring detailed regulatory and technical requirements for the future implementation of our conceptual model and providing a preliminary evaluation of the concept and ideas behind "GenDAI".

2. State of the Art

Bioinformatics software solutions exist to support the analysis of instrument data generated by genomic applications. These solutions can be broadly classified into (i) generic bioinformatics data processing solutions and (ii) application-specific solutions. Generic solutions are developed to support all types of bioinformatics problems and are typically

characterized by a flexible workflow approach, where individual tasks are connected as needed to achieve the intended goal. The flexible workflow concepts can better adapt to the constant progress in science and technology, as individual components can be replaced without altering other components of the system. Their disadvantage is that they are more difficult to set up and use than application-specific standard solutions since the latter can optimize the user experience for specific, relevant use cases.

An example of a generic workflow-based solution is the Galaxy project [15], which has several thousand tools that can be used as tasks in a workflow. It provides a multi-user web interface and is scalable to many concurrent compute nodes. It is available on free public servers, but can also be installed locally. With public servers, there are limitations on the available tools and the maximum amount of resources that can be consumed. Application-specific solutions are inherently less flexible. In the case of metagenomics, these include, for example, MG-RAST [16], MGnify (formerly EBI Metagenomics) [17], and QIIME 2 [18]. A popular and feature-rich solution for gene expression analysis is qBase+ [19,20].

The solutions in both categories have in common that they do not use AI to improve user interaction, e.g., by suggesting appropriate analysis methods or visualizations. Although, in some cases, machine learning algorithms are used in certain analysis steps of these products, this is also rarely enacted [5]. To overcome these and other problems, a conceptual architecture for the specific use case of rumen microbiome analysis was introduced in [6] (Figure 2). It was designed from the beginning to enable the use of AI in all relevant domains. It has a distributed architecture with a workflow engine and task scheduler at its core. To incorporate AI into all aspects of the solution, it was built on the AI2VIS4BigData reference model [21]. Recently, it has also been extended for the use case of human metagenome analysis [5]. As the name implies, AI2VIS4BigData also targets the challenges associated with Big Data processing. While some of the previously mentioned tools, particularly in the context of metagenomics, support the analysis of large data sets through parallel processing and streaming mechanisms, they do not fully address the challenge, as the actual analysis is only one of several steps in a Big Data process.

Another challenge with biomedical software solutions is the difficulty of using most products in laboratory diagnostics due to strict regulatory requirements. Although this use case is explicitly mentioned in the AI2VIS4BigData conceptual architecture for metagenomics, the assessment is very preliminary and does not take into account the applicable standards, such as the ISO standards mentioned above [9–11], and the recent regulation introduced by the IVDR in the European Union [12]. For the use case of gene expression analysis (Figure 3), a model was presented in [22] that is more focused on these regulatory issues and the specific needs of laboratory diagnostics. It is based on the CRISP4BigData reference model [23] (Figure 4). Unfortunately, it was not specifically designed for use with AI and is also limited to the use case of gene expression analysis. To our knowledge, there is no conceptual model that could serve as a template to support large-scale (in terms of Big Data), AI-driven genomic applications specifically tailored to high-throughput laboratory diagnostics.
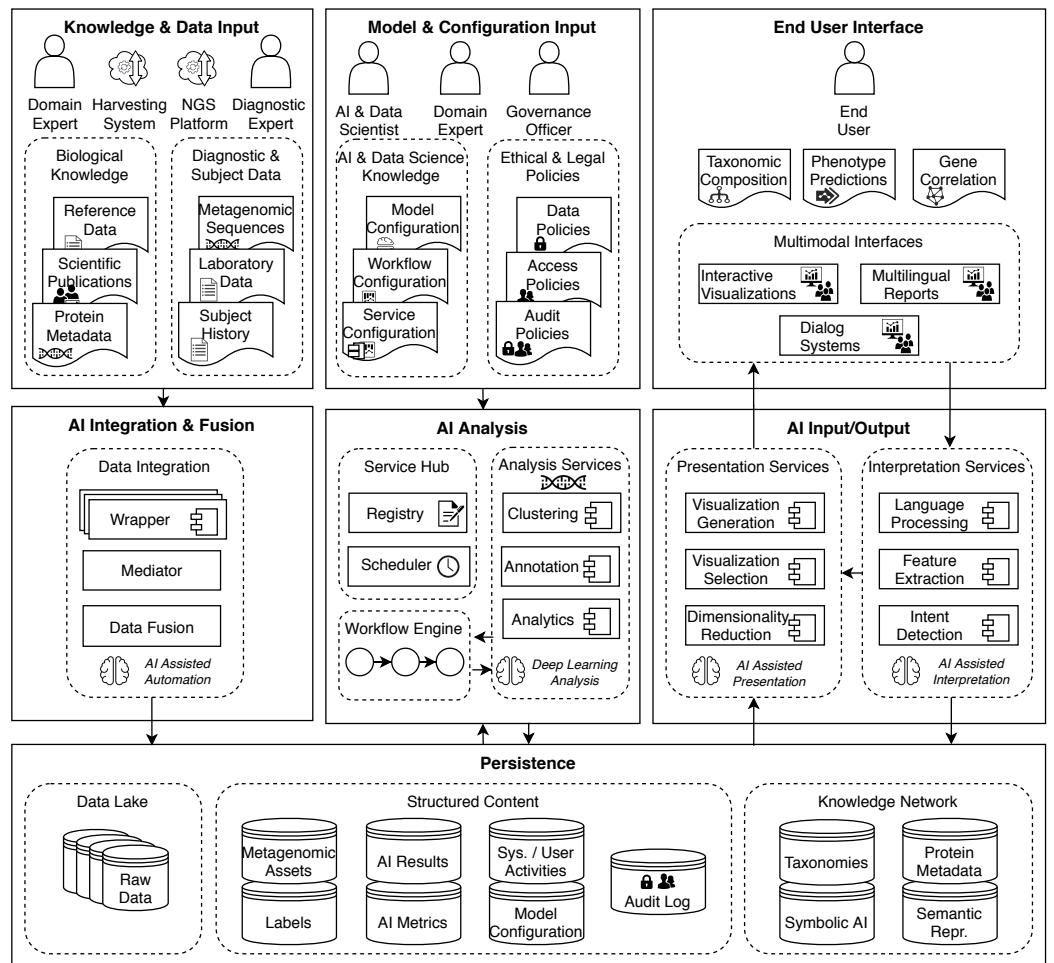
Figure 2. A conceptual architecture for AI and Big Data supporting metagenomics research [6].
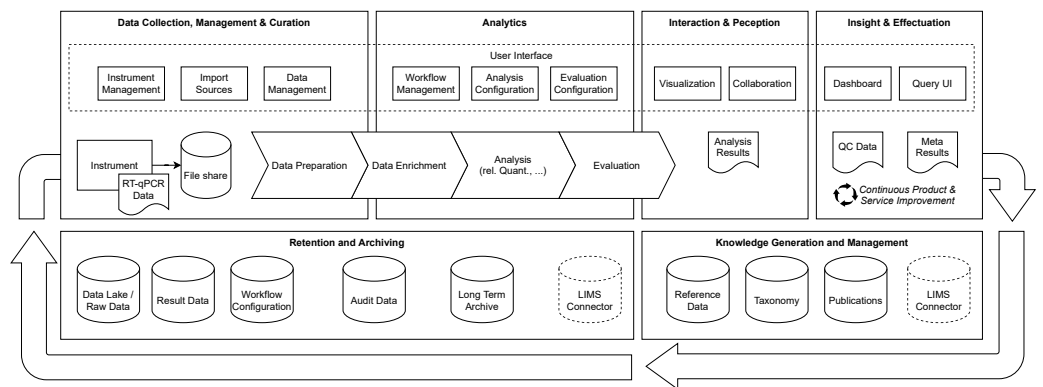

Figure 3. CRISP4BigData-based architecture of gene expression analysis platform [22].
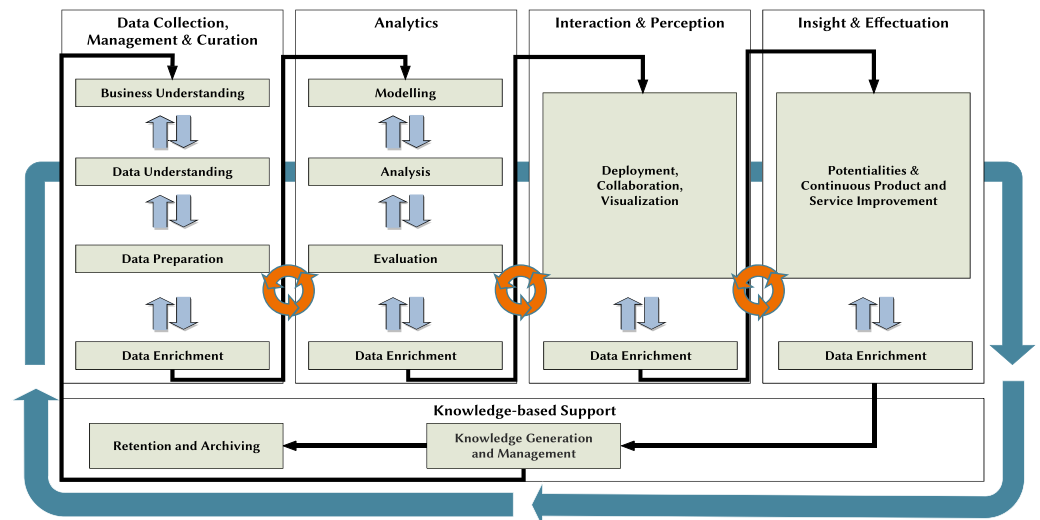
Figure 4. CRISP4BigData reference model [23].

3. Conceptual Model

We propose GenDAI (Figure 5) as a new model that combines the AI-driven nature of the AI2VIS4BigData conceptual architecture for metagenomics with the CRISP4BigData-based model for gene expression diagnostics. GenDAI incorporates elements from both of these earlier models.
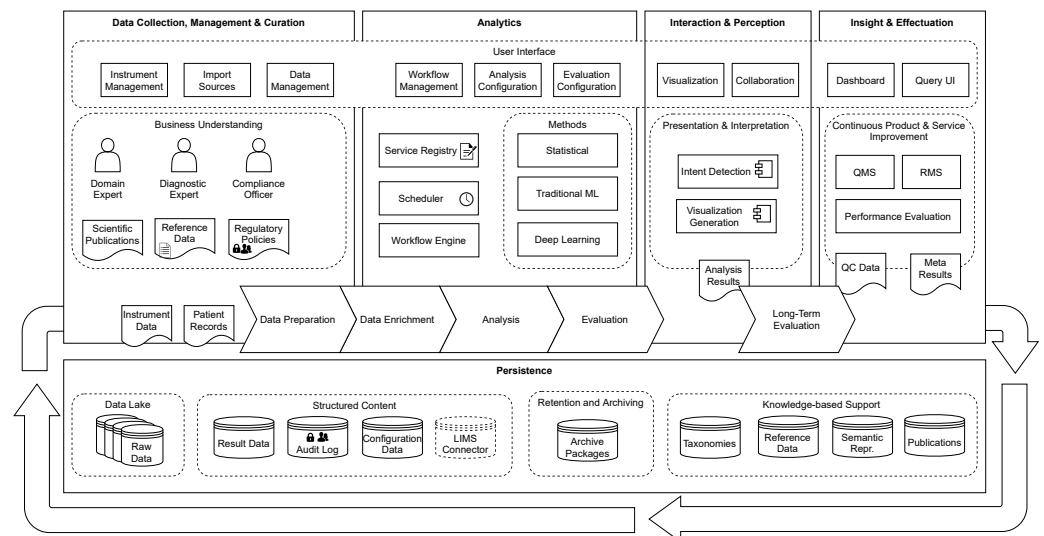
Figure 5. GenDAI conceptual model, © 2021 IEEE. Reprinted, with permission [14].

The general structure of the model is based on the CRISP4BigData reference model, with different phases for "data collection, management, and curation", "analytics", "interaction and perception", and "insight and effectuation". It also follows a three-layer design, with the user interface at the top and a persistence layer at the bottom. In the middle of Figure 5, the data flow through the system is shown as a series of generic processing steps. These steps are mapped by CRISP4BigData and are more granular than their phases.

Unlike the model based on AI2VIS4BigData, there is no explicit, conceptual "AI layer". Instead, the AI components are integrated into their respective phases, simplifying the model in this respect. This change also helps to emphasize the data flow between phases. Even if the specific use of AI is ultimately subject to the respective implementation, the goal of GenDAI is to define interfaces at which AI can be used by clearly naming the individual components and their interaction. Technically, the fundamentally modular approach of

GenDAI supports the flexible replacement of individual components by components with AI support as soon as they have been clinically evaluated and accepted.

Explicit references to specific genomic applications such as metagenomics or gene expression analysis have been removed from the model or generalized so that the model can be used for other applications. The importance of regulatory requirements was enhanced in respect to both models by including (i) the compliance officer as an explicit actor, (ii) regulatory policies as a possible data artifact, and (iii) "long-term evaluation", as an explicit requirement for use in laboratory diagnostics. When analysis is performed using an LDT, part of this long-term evaluation is tracking the LDT within a Quality Management System (QMS), Risk Management System (RMS), and continuous performance evaluation during the entire life cycle of the test. These have been explicitly added as part of the "continuous product and service improvement" topic within the "insight and effectuation" phase.

Looking at the different phases in detail, the first phase, "data collection, management, and curation", concerns all aspects related to data input. In addition to instrument or patient data for analysis, this also includes additional data such as reference data, scientific publications, or applicable policies. These data have been linked to relevant actors and they are considered together as part of the "business understanding" step of CRISP4BigData. This phase also includes the "data preparation" and "data enrichment" steps of CRISP4BigData. However, as a slight deviation from the reference model, the "data enrichment" step has also been pulled into the "analytics" phase, as we believe that with the increasing use of AI and deep learning, data enrichment is often closely related to and dependent on analytics itself. The user interface for this first phase will provide ways to manage import sources, instruments, and (imported) data.

The "analytics" phase includes components required for the actual data analysis, as well as components that manage, organize, and schedule these analytic processes. Examples included in Figure 5 are a workflow engine to orchestrate tasks and data flow, a scheduler to distribute work among compute nodes, and a service registry to manage the list of available tasks and methods. These analysis methods were grouped into three different categories. In addition to statistical methods or classical machine learning approaches, deep learning is included as a separate category to highlight its potential for improved diagnostics. Following the "analysis" step, another important step in the phase is "evaluation". For laboratory diagnostics, it is crucial that the results are checked for plausibility and interpreted. Here, we see potential for future applications of AI to help with both of these challenges.

The third phase, "interaction and perception", concerns the creation of result visualizations. These can be automatically generated reports sent from the lab to the responsible physician, but also visualizations created on demand. For the latter, AI can help to select appropriate visualizations and create them. "Insight and effectuation" in the context of laboratory diagnostics is a phase that focuses on long-term results and meta-analyses rather than single results. Examples include performance evaluation of diagnostic tests performed, as well as risk management systems and quality management systems, which in many cases are required by regulation.

"Persistence" can be considered as both a layer and a phase, because the data are retained and archived after the analysis is complete. However, here, we will consider it a layer because it interacts with all other phases to store intermediate results and can also serve as a data source for initial data import. It should be noted that CRISP4BigData includes a phase called "knowledge-based support", which includes "retention and archiving" as a step, in addition to "knowledge generation and management". We believe that "persistence" is a better term for the heterogeneous types of data managed by the system. However, we retain both steps in the form of data categories in the model. Within the persistence layer are several logical data stores for different data types. The data lake stores, e.g., raw, unprocessed data originating from an instrument. Structured data, on the other hand, may contain, e.g., (interim) results, audit data, or configuration data. A connector for a Laboratory Information Management System (LIMS) is also included, which serves

as an interface to an existing system into which the solution is to be integrated. As a final category, "knowledge-based support" contains knowledge-based data that are independent of individual results. This includes entities such as reference data or taxonomies. They can also be used by AI methods to extract relevant, context-specific information.

In summary, the three core ideas that make up GenDAI are "artificial intelligence" to support end-users in all steps and aspects of the application, "laboratory diagnostics" as a core target market with distinct challenges, and a focus on "genomics" with all its applications.

4. Evaluation and Requirement Engineering

As an initial evaluation of the conceptual model as a valid basis for future implementations, a pre-study was conducted. The pre-study determined the detailed technical and legal requirements of the planned solution. For this, we partnered with a small medical laboratory of ImmBioMed GmbH & Co. KG in Heidelberg, Germany, which provided insights into detailed use cases and processes. The laboratory was selected because it offers various tests for genomic parameters, and the company ImmBioMed also offers consulting services for other laboratories and thus has great experience in the field of laboratory processes. However, as this is only a single laboratory, this evaluation can necessarily only be preliminary. As a practical application for evaluation, we used gene expression analysis of cytokine-dependent genes, which can be an important diagnostic indicator for inflammatory or antiviral defense reactions [24].

Requirements were gathered using a structured approach based on the research framework of Nunamaker et al. [25]. Methods utilized in the approach included a literature review, transcribed interviews, on-site visits, use case modeling, market analysis, and cognitive walkthroughs. A particular focus of the pre-study was the execution of already developed tests as opposed to the development of new tests. In this area, the use cases mapped in Figure 6 were examined in more detail. These use cases can be assigned to the four user stereotypes, "Lab Biologist", "Data Analyst", "Clinical Pathologist", and "QM and Compliance Officer". The evaluation was conducted by matching the identified challenges in the current process and requirements for future solutions with the "GenDAI" model, to determine if and how the model addresses this challenge or requirement. Hence, it is a qualitative approach. Quantification, e.g., in the form of a target benchmark, did not seem appropriate at this time due to the complexity of the various requirements and the early stage of the evaluation. A prototype called "PlateFlow" was used to evaluate user interface concepts with a cognitive walkthrough.

The preliminary study revealed that there are several points in the current laboratory workflow where processes could be more automated. For example, data have to be transferred manually between different systems several times. Due to different data formats and lack of import/export interfaces, this transfer is sometimes conducted by manual entry. Such manual transfer requires special attention and measures, such as a 4-eyes principle to avoid or detect incorrect entries. These, and similar manual steps, cost time and increase throughput times. This confirms the need for GenDAI's holistic approach, where the entire process is mapped and integrated. Table 1 shows an overview of the use cases and an assessment of the automation potential (low/medium/high) in the laboratory studied.

Table 1. Estimated potential for automatization of use cases.

Use Case	Potential			Limitations
	Low	Med.	High	
U1. Prepare Test		x		
U1.1. Program Cycler		x		Cycler Capabilities
U2. Execute Test		x	x	
U2.1. Document Test Protocol		x		User Input
U2.2. Quality Control			x	
U2.3. Store Results			x	

Table 1. *Cont.*

Use Case	Potential			Limitations
	Low	**Med.**	**High**	
U3. Retrieve Results			x	
U3.1. Quality Control			x	
U4. Determine Gene Expression Level			x	
U4.1. Calculate (Δ)ΔCq			x	
U4.2. Apply Formulas			x	
U4.3. Store Analysis Results			x	
U5. Prepare Findings Report		x	x	
U5.1. Summarize Results			x	
U5.2. Summarize interpretation		x		Plausibility Checks
U6. Run Metaanalysis		x		
U6.1. Inter-Run QC		x		Not Formalized
U7. Create Findings Report	x		x	
U7.1. Verify Report	x			Legal Responsibility
U7.2. Submit Report			x	

Figure 6. Use cases analyzed in laboratory.

The interviews also confirmed the notion that laboratories are in an ongoing process of improving existing tests and developing new tests. In addition, there are evolving requirements from the regulatory area. These changes can also have an impact on the IT-supported processes, which is why the software and systems used need to be flexible or adaptable enough to meet changing requirements. In GenDAI, this need is underscored by the use of individual interchangeable and extensible components and by a flexible overarching workflow.

The market analysis revealed that most of the tools were outdated or did not offer all the analysis and processing functions needed. Moreover, none of the tools examined were specifically designed to meet the needs of medical laboratories. For example, the tools did not meet the necessary regulatory requirements, did not cover the complete workflow, and their user interface was also designed more for scientific research or the development of new tests, rather than the efficient processing of tests already developed. This also confirms the need for a new solution specifically designed for laboratory processes. Table 2 shows an overview of the different software tools for gene expression analysis that were evaluated, including their basic functionalities and last update date. Table 2 is a summary of the results given in Krause et al. [26], which were, in turn, based on the results of Pabinger et al. [27]. A "+" symbolizes the presence of a feature, a "+" the absence. Features whose existence could not be reliably determined have been marked as "nd" (not determined).

Table 2. qPCR software evaluation. Summarized from Krause et al. [26], Pabinger et al. [27].

Tool	PCR Efficiency Estimation	Melt Curve Analysis	Reference Gene Selection	Cq Calculation	Error Propagation	Normalization	Absolute Quantifcation	Relative Quantification	Outlier Detection	NA Handling	Statistics	Graphs	MIQE Compliant	Last Update
CAmpER	+	nd	nd	+	−	−	−	+	nd	−	−	+	−	2009
Cy0 Method	−	−	−	+	−	−	−	−	−	−	−	−	+	2010
DART-PCR	−	−	−	+	−	+	−	+	+	−	−	+	−	2002
Deconvolution	−	−	−	−	−	−	+	−	−	−	−	−	+	2010
ExpressionSuite Software	−	+	−	+	−	+	−	+	+	−	+	+	+	2019
Factor-qPCR	−	−	−	−	−	+	−	−	−	−	−	−	+	2020
GenEx	+	−	+	−	−	+	+	+	+	+	+	+	+	2019
geNorm	−	−	+	−	−	−	−	−	−	−	−	−	−	2018
LinRegPCR	+	−	−	+	−	−	+	−	+	−	−	+	+	2021
LRE Analysis	−	−	−	−	−	−	+	−	−	−	−	−	+	2012
LRE Analyzer	−	−	−	−	−	−	+	−	−	−	−	+	+	2014
MAKERGAUL	−	−	−	+	−	−	+	−	−	−	−	−	+	2013
PCR-Miner	+	−	−	+	−	−	−	−	−	−	−	−	+	2011
PIPE-T	−	−	−	−	−	+	+	+	+	+	+	+	−	2019
pyQPCR	+	−	−	−	+	+	−	+	−	+	−	+	+	2012
Q-Gene	+	−	−	−	−	+	−	+	−	−	−	+	−	2002
qBase	+	−	+	−	+	+	−	+	+	−	+	+	+	2007
qbase+	+	−	+	−	+	+	+	+	+	−	+	+	+	2017
qCalculator	+	−	−	−	−	+	−	+	−	+	−	+	−	2004
QPCR	+	−	−	+	+	+	−	+	−	+	+	+	+	2013
qPCR-DAMS	−	−	−	−	−	+	+	+	−	+	−	−	+	2006
RealTime StatMiner	−	−	+	−	+	+	−	+	+	+	+	+	+	2014
REST	−	−	−	−	+	+	−	+	−	−	+	+	+	2009
SARS	−	nd	nd	−	−	+	−	+	nd	−	+	−	+	2011
SoFAR	+	+	−	+	−	−	−	−	−	−	−	+	−	2003

In order to validate possible operating concepts for a solution, the PlateFlow proof-of-concept prototype was developed as part of the pre-study, which allows relevant analyses to be performed from the raw data and the results to be summarized in a report. Figure 7 shows one of the screens of PlateFlow. PlateFlow was evaluated through a cognitive walkthrough, which resulted in positive feedback regarding the scope of functions. In addition, the need to further evaluate usability in future development was highlighted.
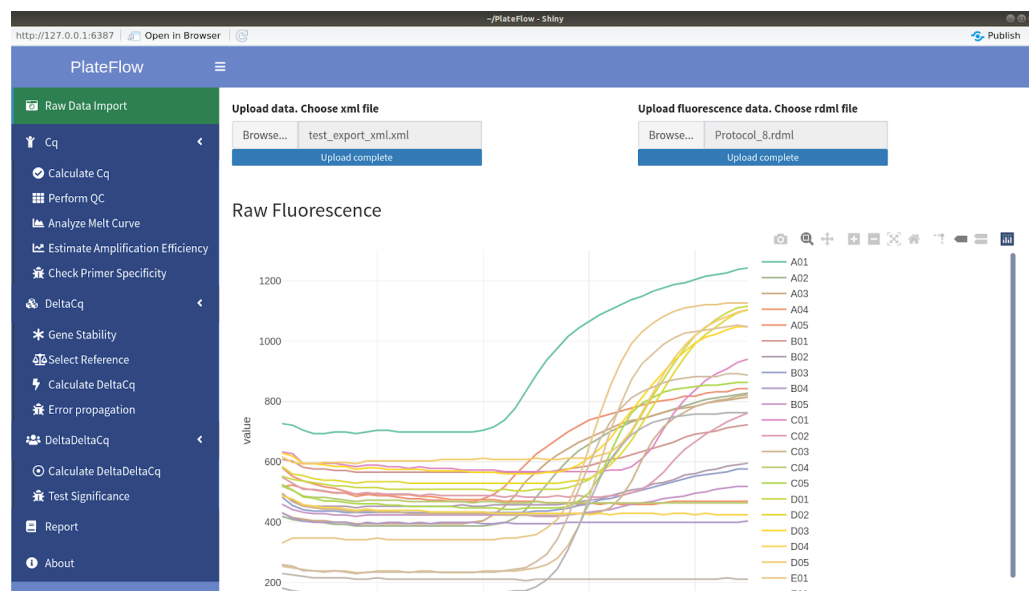

Figure 7. "PlateFlow" prototype user interface.

5. Conclusions and Future Work

The use of AI-assisted genomic applications has the potential to improve laboratory diagnostics. However, regulatory and other challenges currently hinder greater innovation in this area. There is a need for a software platform for genomic applications in laboratory diagnostics that leverages AI whilst providing the necessary foundation for regulatory compliance.

Here, we have presented GenDAI as one possible solution. It combines the knowledge of previous architectural models developed for specific genomic applications in different focus areas. Unlike these previous models, GenDAI is independent of specific genomic applications. Unlike the AI2VIS4BigData-based model, it considers the specific requirements of laboratory diagnostics to a much greater extent. Unlike the CRISP4BigData-based model, the integration of AI is also an essential feature of the model. GenDAI thus represents an improvement over both models.

A pre-study in cooperation with a small laboratory enabled a first practical evaluation of the concepts. Part of the preliminary study included the creation of use cases, evaluation of existing software components, requirement engineering, and development of the PlateFlow prototype. The remaining challenges include further practical validation of the model for additional use cases and in other laboratories, a technical architecture, implementation of missing components, and, ultimately, certification of the solution for clinical diagnostics.

Author Contributions: Conceptualization, T.K. and M.H.; investigation, T.K. and E.J.; writing—original draft preparation, T.K. and E.J.; writing—review and editing, S.B., P.M.K., M.K. and M.H.; visualization, T.K. and E.J.; supervision, M.K. and M.H.; project administration, T.K. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stephens, Z.D.; Lee, S.Y.; Faghri, F.; Campbell, R.H.; Zhai, C.; Efron, M.J.; Iyer, R.; Schatz, M.C.; Sinha, S.; Robinson, G.E. Big Data: Astronomical or Genomical? *PLoS Biol.* **2015**, *13*, e1002195,
2. Abawajy, J. Comprehensive analysis of big data variety landscape. *Int. J. Parallel Emergent Distrib. Syst.* **2015**, *30*, 5–14. [CrossRef]
3. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
4. Zou, J.; Huss, M.; Abid, A.; Mohammadi, P.; Torkamani, A.; Telenti, A. A primer on deep learning in genomics. *Nat. Genet.* **2019**, *51*, 12–18.
5. Krause, T.; Wassan, J.T.; Mc Kevitt, P.; Wang, H.; Zheng, H.; Hemmje, M.L. Analyzing Large Microbiome Datasets Using Machine Learning and Big Data. *BioMedInformatics* **2021**, *1*, 138–165. [CrossRef]
6. Reis, T.; Krause, T.; Bornschlegl, M.X.; Hemmje, M.L. A Conceptual Architecture for AI-based Big Data Analysis and Visualization Supporting Metagenomics Research. In Proceedings of the Collaborative European Research Conference (CERC 2020), Belfast, UK, 10–11 September 2020; Afli, H., Bleimann, U., Burkhardt, D., Loew, R., Regier, S., Stengel, I., Wang, H., Zheng, H., Eds.; CEUR Workshop Proceedings; CERC: New Delhi, India, 2020; pp. 264–272.
7. Soueidan, H.; Nikolski, M. Machine learning for metagenomics: Methods and tools. *arXiv* **2015**, arXiv:1510.06621.
8. Hamon, R.; Junklewitz, H.; Sanchez, I. *Robustness and Explainability of Artificial Intelligence*; EUR, Publications Office of the European Union: Luxembourg, 2020; Volume 30040.
9. *Standard ISO 13485:2016*; Medical Devices—Quality Management Systems—Requirements for Regulatory Purposes. ISO International Organization for Standardization: Geneva, Switzerland, 2016.
10. *Standard ISO 15189:2012*; Medical Laboratories—Requirements for Quality and Competence. ISO International Organization for Standardization: Geneva, Switzerland, 2012.
11. *Standard IEC 62304:2006*; Medical Device Software—Software Life Cycle Processes. IEC International Electrotechnical Commission: Geneva, Switzerland, 2006.
12. The European Parliament; The Council of the European Union. *In Vitro Diagnostic Regulation*; European Commission: Brussels, Belgium, 2017.
13. Spitzenberger, F.; Patel, J.; Gebuhr, I.; Kruttwig, K.; Safi, A.; Meisel, C. Laboratory-Developed Tests: Design of a Regulatory Strategy in Compliance with the International State-of-the-Art and the Regulation (EU) 2017/746 (EU IVDR In Vitro Diagnostic Medical Device Regulation). *Ther. Innov. Regul. Sci.* **2021**, *56*, 47–64.
14. Krause, T.; Jolkver, E.; Bruchhaus, S.; Kramer, M.; Hemmje, M.L. GenDAI—AI-Assisted Laboratory Diagnostics for Genomic Applications. In Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 9–12 December 2021. [CrossRef]
15. Afgan, E.; Baker, D.; Batut, B.; van den Beek, M.; Bouvier, D.; Čech, M.; Chilton, J.; Clements, D.; Coraor, N.; Grüning, B.A.; et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **2018**, *46*, W537–W544. [CrossRef] [PubMed]
16. Meyer, F.; Paarmann, D.; D'Souza, M.; Olson, R.; Glass, E.M.; Kubal, M.; Paczian, T.; Rodriguez, A.; Stevens, R.; Wilke, A.; et al. The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* **2008**, *9*, 1–8. [CrossRef] [PubMed]
17. Mitchell, A.L.; Almeida, A.; Beracochea, M.; Boland, M.; Burgin, J.; Cochrane, G.; Crusoe, M.R.; Kale, V.; Potter, S.C.; Richardson, L.J.; et al. MGnify: The microbiome analysis resource in 2020. *Nucleic Acids Res.* **2020**, *48*, D570–D578. [CrossRef] [PubMed]
18. Bolyen, E.; Rideout, J.R.; Dillon, M.R.; Bokulich, N.A.; Abnet, C.C.; Al-Ghalith, G.A.; Alexander, H.; Alm, E.J.; Arumugam, M.; Asnicar, F.; et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **2019**, *37*, 852–857.
19. What Makes qbase+ Unique? Available online: https://www.qbaseplus.com/features (accessed on 7 June 2021).
20. Hellemans, J.; Mortier, G.; de Paepe, A.; Speleman, F.; Vandesompele, J. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol.* **2007**, *8*, R19,
21. Reis, T.; Bornschlegl, M.X.; Hemmje, M.L. Toward a Reference Model for Artificial Intelligence Supporting Big Data Analysis. In *Advances in Data Science and Information Engineering*; Stahlbock, R., Weiss, G.M., Abou-Nasr, M., Yang, C.Y., Arabnia, H.R., Deligiannidis, L., Eds.; Transactions on Computational Science and Computational Intelligence; Springer: Cham, Switzerland, 2021._38. [CrossRef]
22. Krause, T.; Jolkver, E.; Bruchhaus, S.; Kramer, M.; Hemmje, M.L. An RT-qPCR Data Analysis Platform. In Proceedings of the Collaborative European Research Conference (CERC 2021), Cork, Ireland, 9–10 September 2021; Afli, H., Bleimann, U., Burkhardt, D., Hasanuzzaman, M., Loew, R., Reichel, D., Wang, H., Zheng, H., Eds.; CEUR Workshop Proceedings; CERC: New Delhi, India, 2021.

23. Berwind, K.; Bornschlegl, M.X.; Kaufmann, M.A.; Hemmje, M.L. Towards a Cross Industry Standard Process to support Big Data Applications in Virtual Research Environments. In Proceedings of the Collaborative European Research Conference (CERC 2016), Cork, Ireland, 23–24 September 2016; Bleimann, U., Humm, B., Loew, R., Stengel, I., Walsh, P., Eds.; CERC: New Delhi, India, 2016.

24. Barrat, F.J.; Crow, M.K.; Ivashkiv, L.B. Interferon target-gene expression and epigenomic signatures in health and disease. *Nat. Immunol.* **2019**, *20*, 1574–1583.

25. Nunamaker, J.F.; Chen, M.; Purdin, T.D. Systems Development in Information Systems Research. *J. Manag. Inf. Syst.* **1990**, *7*, 89–106. [CrossRef]

26. Krause, T.; Jolkver, E.; Mc Kevitt, P.; Kramer, M.; Hemmje, M. A Systematic Approach to Diagnostic Laboratory Software Requirements Analysis. *Bioengineering* **2022**, *9*, 144.

27. Pabinger, S.; Rödiger, S.; Kriegner, A.; Vierlinger, K.; Weinhäusel, A. A survey of tools for the analysis of quantitative PCR (qPCR) data. *Biomol. Detect. Quantif.* **2014**, *1*, 23–33.