Bilkent University

Department of Computer Engineering

Senior Design Project

AnnoHub

Analysis & Requirements Report

T2331

Emirkan Derköken, Bora Yılmaz, Dağhan Ünal, İlker Özgen, Onurcan Ataç

Supervisor: Shervin R. Arashloo

Table of Contents

1. Introduction

In the realm of machine learning and artificial intelligence, the demand for accurately labeled data faces formidable challenges. The necessity for precise annotations clashes with the complexities of acquiring and maintaining a skilled workforce for data labeling tasks. This dilemma poses a significant hurdle for businesses seeking reliable datasets.

AnnoHub emerges as a solution designed to tackle these challenges head-on. This innovative platform offers a dual-mode operation catering to both labelers and companies. It enables collaborative annotations through freelance annotators and facilitates private projects, effectively bridging the gap between data demand and workforce availability.

More than just an operational tool, AnnoHub integrates sophisticated features to uphold annotation integrity. Through validation techniques, error-reducing machine learning algorithms, and robust fraud detection, the platform aims to elevate annotation quality and reliability.

2. Problem Definition

Accurate model construction in the rapidly developing fields of artificial intelligence and machine learning requires high-quality labeled data. Large-scale data labeling, however, presents difficulties, especially when it comes to guaranteeing uniformity and accuracy. This is crucial since it has an immediate effect on the accuracy and resilience of produced models. For large datasets, obtaining a consistent and trustworthy data annotation requires careful approaches and strict quality control. Furthermore, managing sensitive data increases complexity and necessitates striking a careful balance between protecting individual privacy and company secrets and annotating the dataset.

Beyond the technical intricacies, sourcing a skilled workforce for data annotation poses a fundamental challenge, requiring careful consideration from recruitment to ongoing training. In contrast to the nuanced challenges of acquiring skilled labor for complex tasks, finding unskilled workers for simple data annotation is also difficult due to the sporadic nature of the work. Even though the machine learning industry at large exhibits a high demand for labeled data; individual companies encounter intermittent demands, characterized by urgent requirements during development phases. These spikes in demand are unpredictable and occur sporadically throughout the life-cycle of a business, with periods of intense need followed by times when data labeling is not required. As a result; companies are reluctant to establish in-house teams, given the irregular demand, making it economically impractical. Simultaneously, potential workers are deterred by the lack of stability in roles, resulting in a scarcity of individuals willing to perform straightforward data annotation tasks.

With its intuitive interfaces, AnnoHub is intended to provide a high-quality user experience while addressing the primary issues associated with data annotation, as previously discussed. Rich in features that improve the lives of annotators and businesses alike, guaranteeing a productive and successful annotation process.

3. Proposed Systems

3.1 System Overview

AnnoHub is a platform for thorough data annotation that supports two different modes of operation. Users have the option to openly share unlabeled datasets and hire independent annotators to work together to annotate them. Compared to engaging in-house annotators, this operational mode relieves publishers of the effort of finding the required manpower and annotates their data at a more economical rate. Furthermore, this alternative offers freelance annotators a new stream of side income that practically everyone may tap into. Alternatively, companies have the option to create private data annotation projects with their datasets, utilizing the platform as a sophisticated labeling tool to enlist annotators of their choosing.

AnnoHub integrates certain features to improve the consistency and accuracy of annotations, especially those made by external annotators on publicly available datasets, in order to look for potential contradictions both between and among annotators. These include methods for verifying the accuracy of the annotators' work, like integrating known-annotation trick data into the dataset, self-validation, which requires annotators to label the same data multiple times, and cross-validation, which involves the same data being annotated by multiple annotators. Furthermore; through the utilization of unsupervised machine learning algorithms for mislabel prediction, the labeling error rate is aimed to be reduced. Simultaneously; by the implementation of fraud detection algorithms, the negative effects of malicious use are aimed to be minimized. These measures and the information they are going to present to the companies collectively enhance the reliability and quality assurance of the annotation process on our platform.

Finally, owing to the AnnoHub's system architecture prioritizing enhanced security, in cases where companies choose to create a private data annotation project, the data is going to bypass our servers entirely. Instead, it is directly shared with the annotators, ensuring the safeguarding of company secrets and adherence to data protection laws not through bonds, but through the system design.

3.2 Functional Requirements

3.2.1 Data Ingestion and Partitioning

- Support for various data types including text, images, audio, video, and 3D point clouds. More can be added in the future.
- Data partitioning to ensure privacy and manageability. If chosen so, data can also be partitioned manually by clients, where the data is sensitive even in smaller subsets. Otherwise, automatically partition data.
- Clients can choose to store their data directly to our servers or hide it so that the data can be sent to the annotators directly bypassing our servers while still using our front-end and back-end logic.

3.2.2 Data Labeling and Annotation

- User-friendly interface for labeling data.
- Support for various labeling tasks such as classification, object detection, segmentation, etc.
- Classification algorithms that help identify the category or class of an object within an image.
- Object detection algorithms that help identifying and locating the objects within an image by drawing a box around each detected object.
- Semantic segmentation process that labels each pixel in an image with a class.
- Instance segmentation process that differentiates between different instances of the same class such as two different types of cars.

3.2.3 Quality Assurance

- Automated and manual quality checks to ensure label accuracy
- Real-time monitoring and reporting of labeling accuracy and labeler performance.
- Ability for supervisors to review and correct labels.
- Ability for supervisors to rate labelers based on their performance.
- Qualification tests and exams that make sure each labeler is qualified for labeling the data provided. Different tests can be provided from the clients based on the requirements of the data.
- Fraud detection algorithms that dig through annotators to detect any frauds and tag labelers in the danger zone for further view.
- Cross validation and self-validation techniques that train the model in some subsets and test it on the others.
- Mislabel prediction using unsupervised machine learning algorithms that aims to identify possibly mislabeled data in datasets.
- Usage of trick data to determine false labelers to further validate their labels.

3.2.4 Data Aggregation and Retrieval

- Tools to aggregate labeled data from various labeling tasks, enabling an organized collection of labeled datasets for further use.
- Efficient data retrieval mechanisms to allow quick access to the labeled data, aiding in timely analysis and usage.
- Indexed data storage to expedite the search and retrieval of specific data sets or data points and features to export aggregated data in various formats for further analysis or for sharing with stakeholders.
- Ability to run queries to filter and retrieve specific subsets of data based on certain criteria and filters.

3.2.5 User Management and Access Control

- Role-based access control to define permissions for different user roles.
- Secure authentication and authorization mechanisms to ensure data privacy
- Trackable user actions to ensure accountability.

3.2.6 Reporting and Analytics

- Dashboards to monitor project progress, labeler performance, and data quality.
- Exportable reports for sharing with stakeholders.
- Analytics tools to derive insights from labeling data and improve project outcomes.

3.2.7 Notification and Communication

- Automated notifications for task assignments, completions, and quality issues, communication tools for collaboration and discussion among labelers and supervisors.
- Feedback mechanism for labelers to report issues/bugs and suggest improvements.

3.3 Non-Functional Requirements

3.3.1 Scalability and Performance

- Ability to handle and optimize increasing volumes of data without compromising performance.
- Scalable workforce management to scale labeling capacity up or down based on demand.
- Optimization tools to further enhance performance.

3.3.2 Data Privacy and Security

- Encryption of sensitive data at rest and in transit, data masking and anonymization tools to protect privacy.
- Taking industry standards, state laws and regulations into account regarding data privacy and security.

3.3.3 System Availability and Reliability

- Reliable data storage and backup solutions to prevent data loss.
- Error handling and recovery mechanisms to ensure system stability and reliability.

3.3.4 Extensibility and Modularity

- Modular system design to allow for the addition of new features and components more easily.
- Extensibility to easily integrate with other systems and tools.
- API documentation for developers to extend and integrate the system.

3.3.5 Interoperability [1]

- Standards-compliant data formats to ensure interoperability with other systems.
- SDKs for integration with external systems and data sources.
- Compatibility with common data labeling and machine learning tools.

3.3.6 Testability

- Testing tools and frameworks to validate system functionality, stability, reliability and performance.

[1] One important thing to note is that in the context of data labeling, annotation format is crucial. It specifies how the labels or annotations are structured and associated with the data being labeled. For instance, in object detection tasks, annotations might include coordinates of bounding boxes and class labels.

3.4 Pseudo Requirements

The choice of React for the frontend aligns with its extensive ecosystem of third-party libraries, fostering rapid development and scalability. React's component-based architecture allows for modular development, enabling efficient management of complex user interfaces. This choice empowers seamless integration and swift iteration, crucial in delivering a responsive and dynamic user experience.

Python stands as the backend language of choice primarily for its unparalleled agility, emphasizing rapid development and accelerated implementation speeds. Its concise syntax facilitates swift prototyping and efficient iteration and leveraging Python's succinctness and rich ecosystem of libraries streamlines the development process, ensuring faster deployment. While Python's run-time speed might appear slower compared to some other choices, the computational focus of AnnoHub primarily resides in its machine learning algorithms. Python serves as an efficient wrapper for these algorithms, leveraging highly optimized C code within its libraries. This architecture positions Python as a frontrunner for running machine learning operations, surpassing competitors due to its adeptness in executing these algorithms swiftly and efficiently. Finally, the backend's primary role involves data read-write operations with the database, rather than intense data computations. This underscores the paramount importance of implementation speed over run-time speed for AnnoHub. In this context, the marginal differences in run-time speed among various backend environments become inconsequential, as the backend's efficiency in swiftly executing operations and handling data transactions holds far more significance in the system's overall performance.

Employing PostgreSQL for user data management leverages its reliability, transactional support, and ACID compliance, ideal for maintaining consistent user information and access control. MongoDB complements this architecture by storing labeling data in a NoSQL format, offering flexibility in handling unstructured or semi-structured data. Its schema-less nature suits the variable nature of labeling data, enhancing efficiency and scalability in this specific domain, which might be inefficient to handle within a rigid relational model.

The decision to utilize Amazon Web Services (AWS) as the cloud provider is rooted in its robust infrastructure, scalability, and comprehensive suite of services. AWS offers a

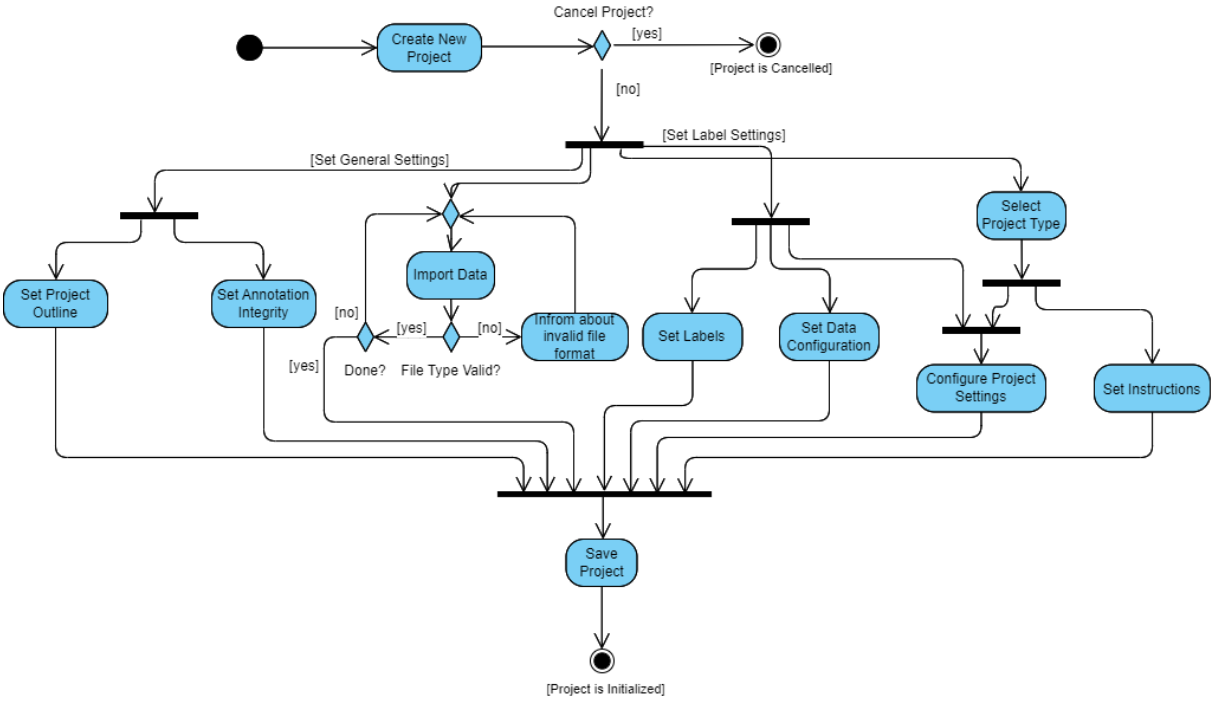
scalable and secure environment, providing a wide array of tools and services tailored to accommodate the varying needs of a complex system like AnnoHub. Its reliability, security features, and global accessibility align with the demands of a platform handling sensitive data and requiring high availability and resilience.

3.5 System Model

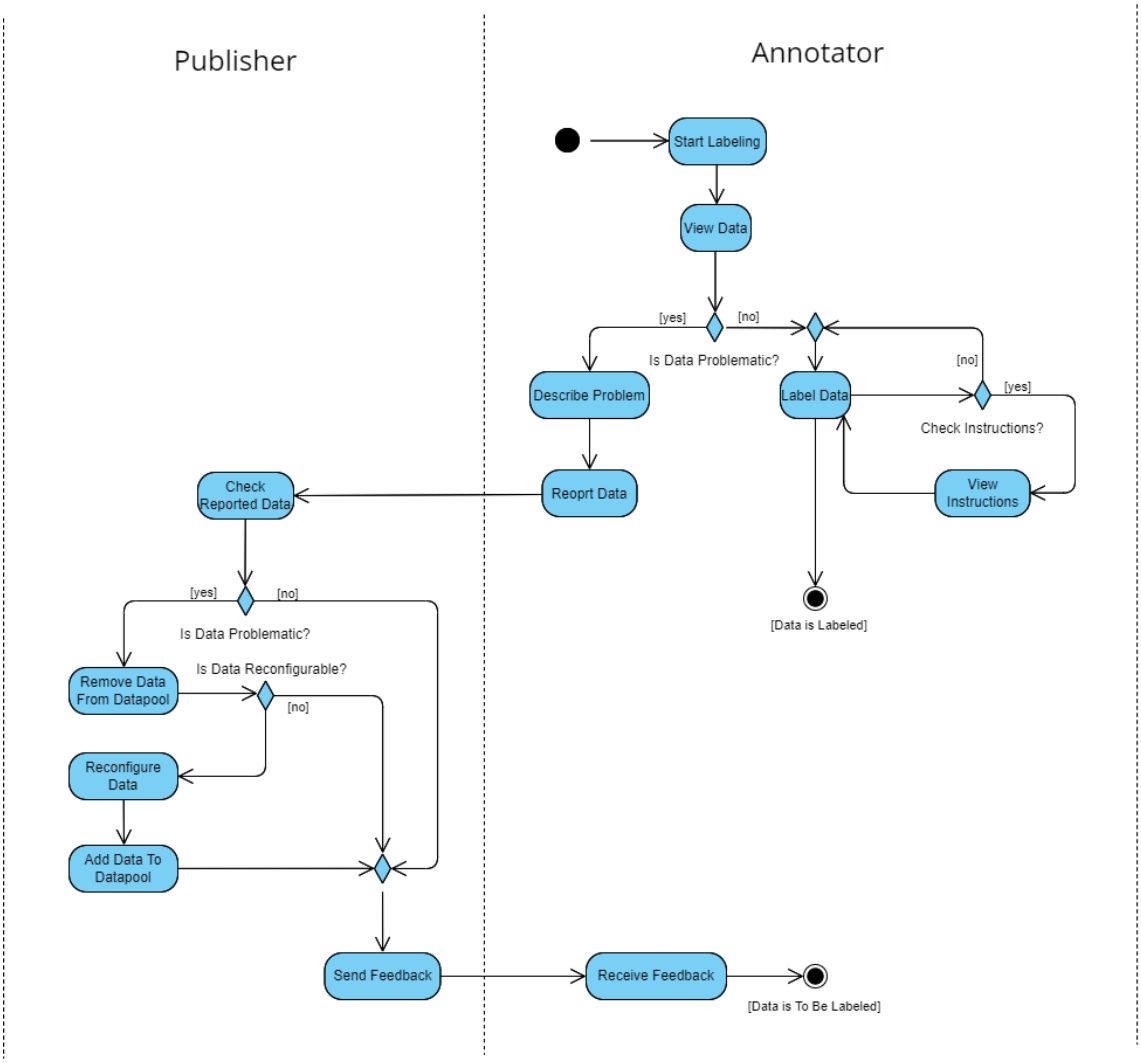
3.5.1 Use Case Diagram

3.5.2 Activity Diagrams

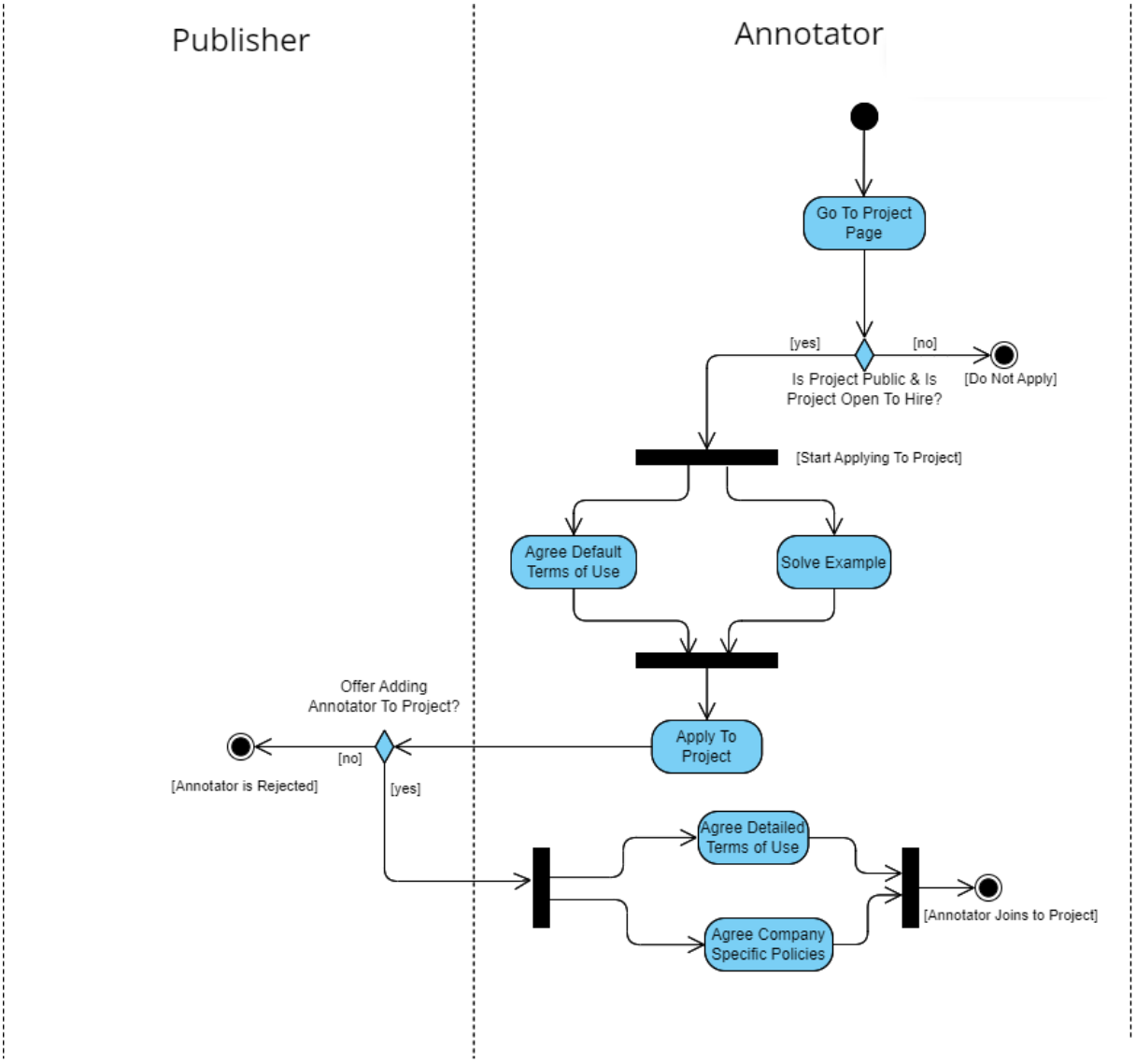
3.5.2.1 Project Creation Activity Diagram



3.5.2.2 Data Labeling Activity Diagram

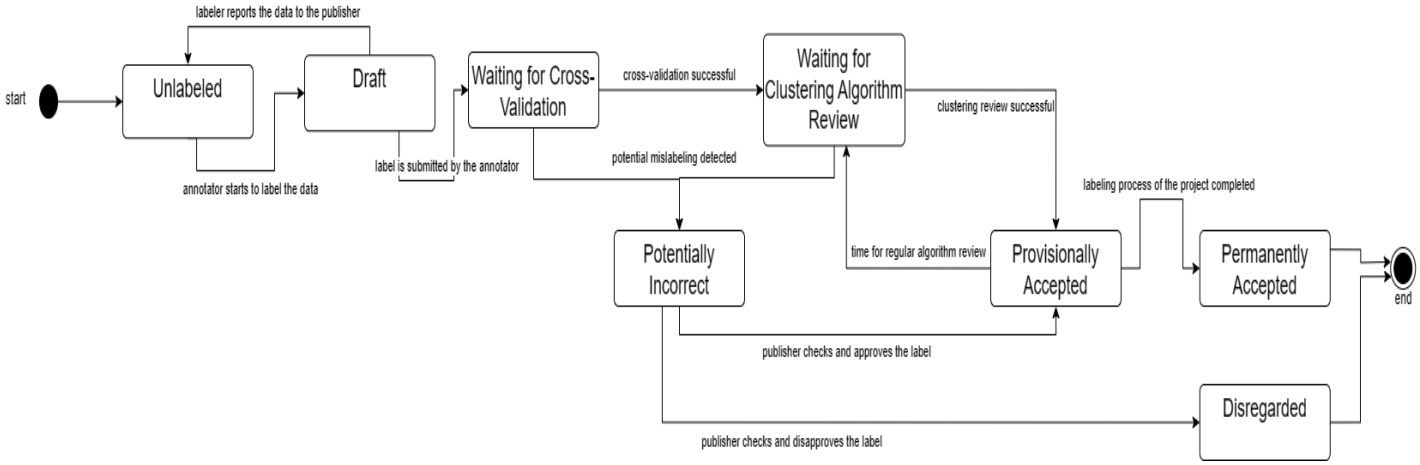


3.5.2.3 Apply Project Activity Diagram

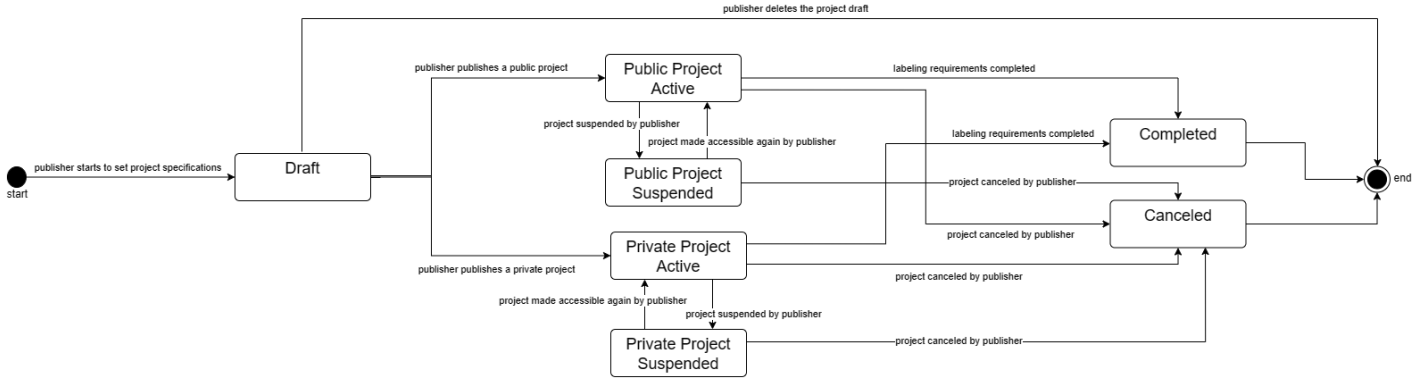


3.5.3 State Diagrams

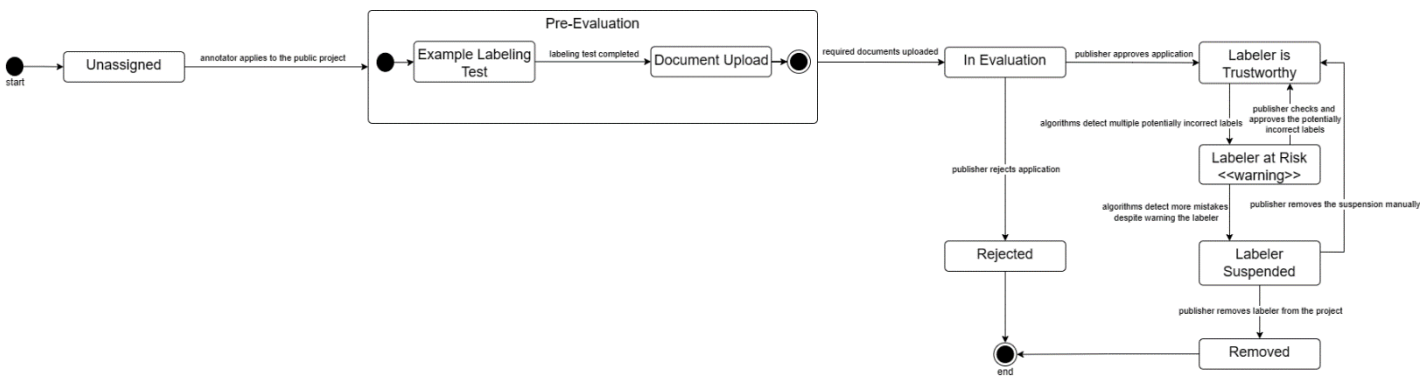
3.5.3.1 Label Instance State Diagram



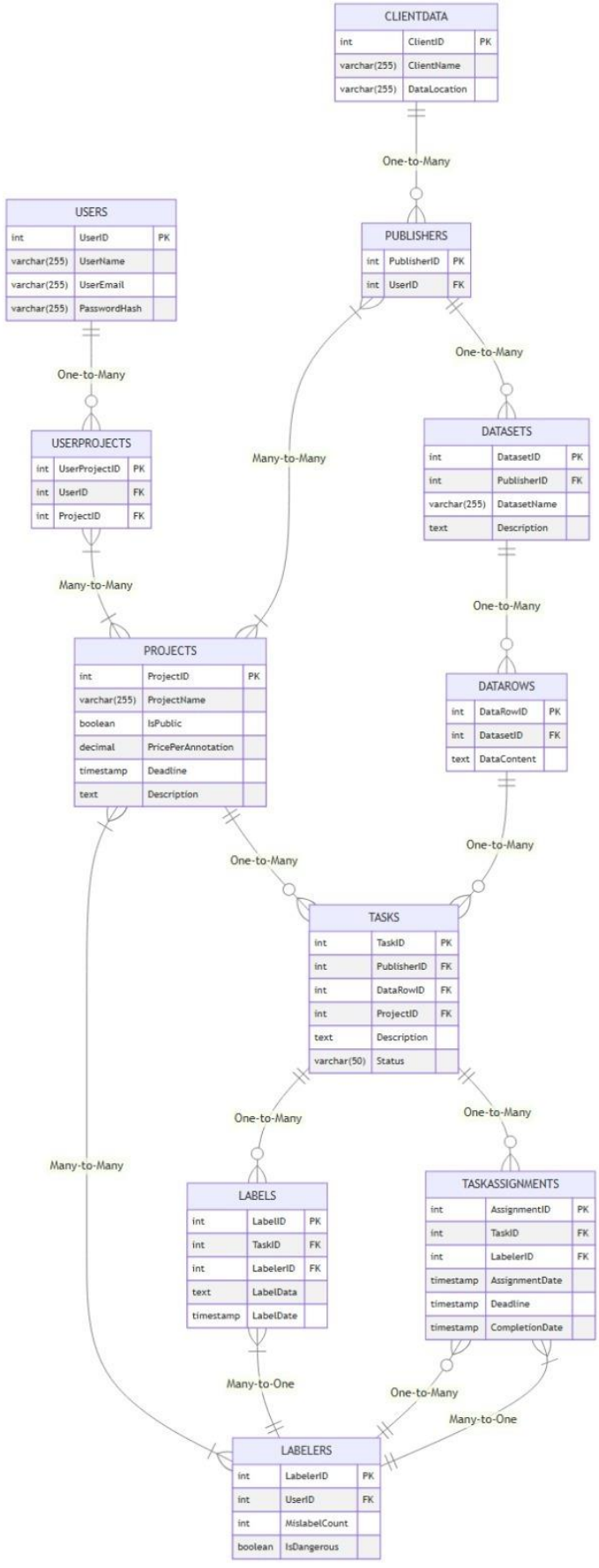
3.5.3.2 Project State Diagram



3.5.3.3 Annotator State Diagram

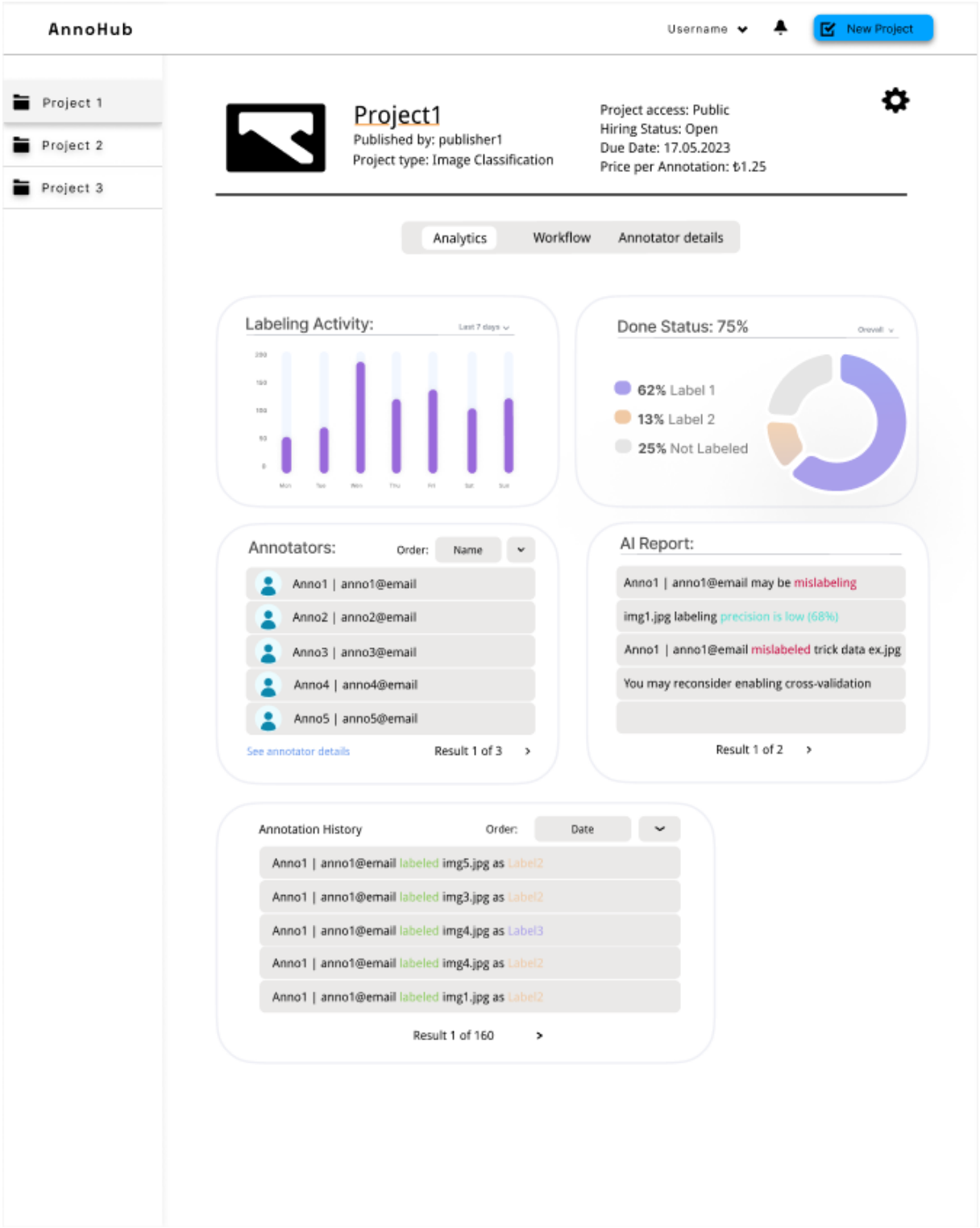


3.5.4 Database Design



3.5.5 User Interfaces

3.5.5.1 Publisher Project Overview



AnnoHub

Username

New Project

- Project 1
- Project 2
- Project 3

Project1

Published by: publisher1
Project type: Image Classification

Project access: Public
Hiring Status: Open
Due Date: 17.05.2023
Price per Annotation: ৳1.25

Analytics | Workflow | Annotator details

all (4)
to label (3)
labeled (2)
open issues (1)
done (1)

ID	Workflow Status	Dataset	Annotator	Open Issues	Precision	
aasgdsaf.jpg	labeled	shelter	Anno1, Anno2	No	99%	✖
abbfdsbc.jpg	labeled	shelter	Anno2, Anno3	Yes	42%	✖
adbfdsbd.png	to label	shelter	-	-	-	✖
adbfdsbe.png	to label	house	-	-	-	✖
adbfdsbf.png	to label	school	-	-	-	✖

Result 1 of 1

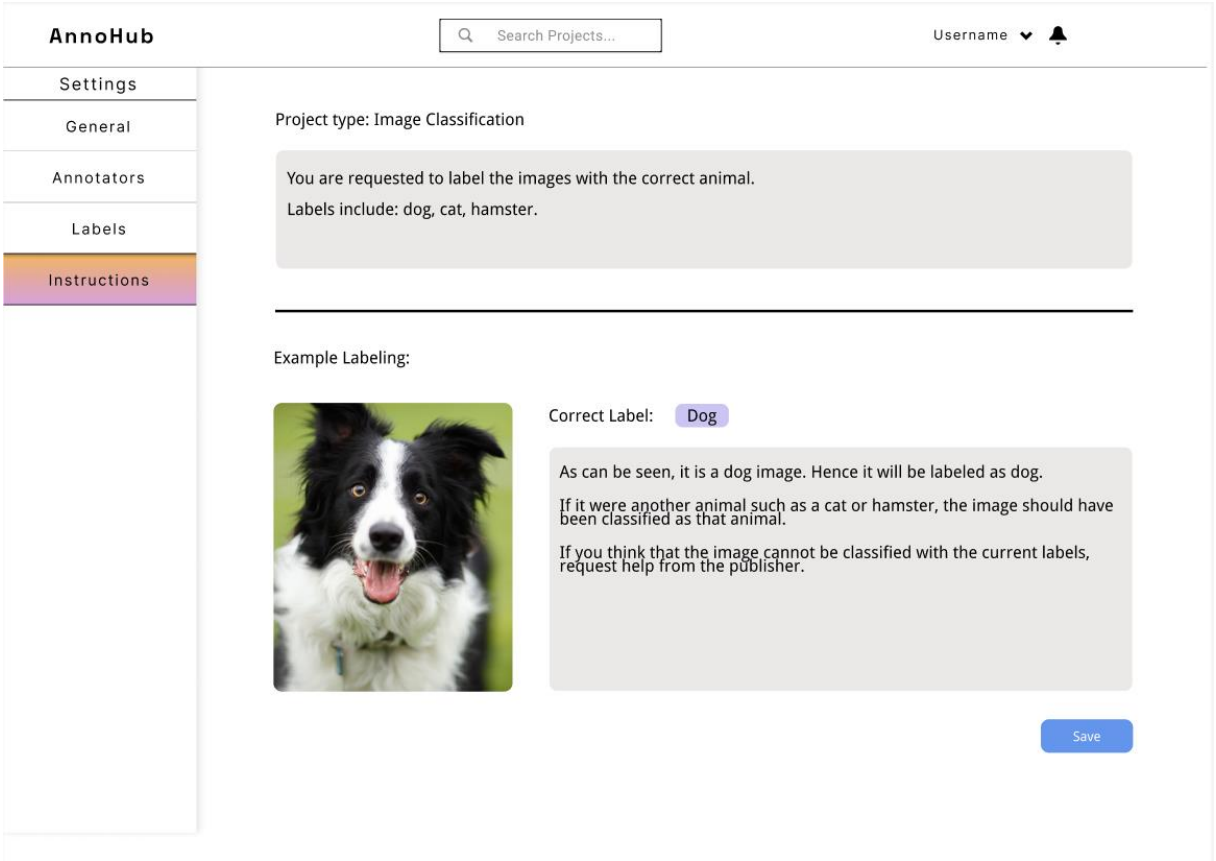
Upload Supporting Files

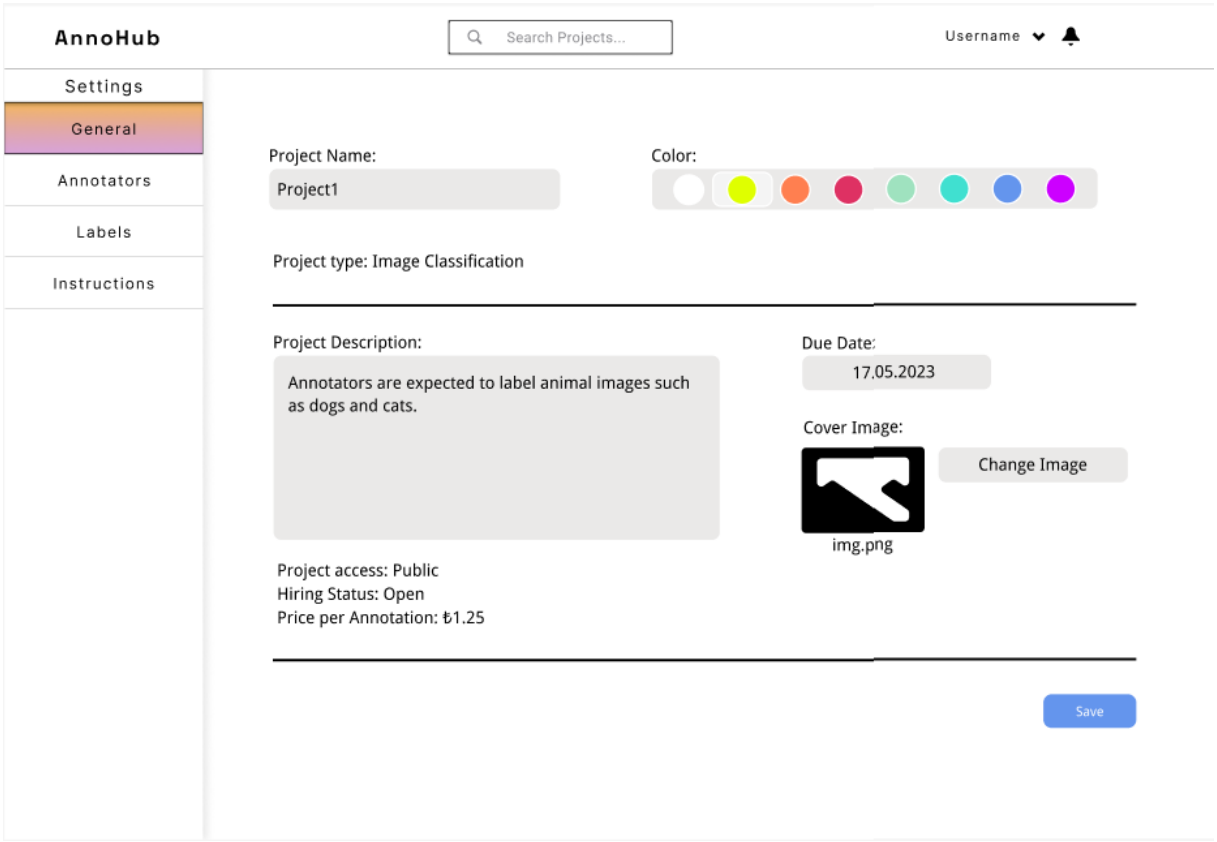
Choose files to upload | Browse Files

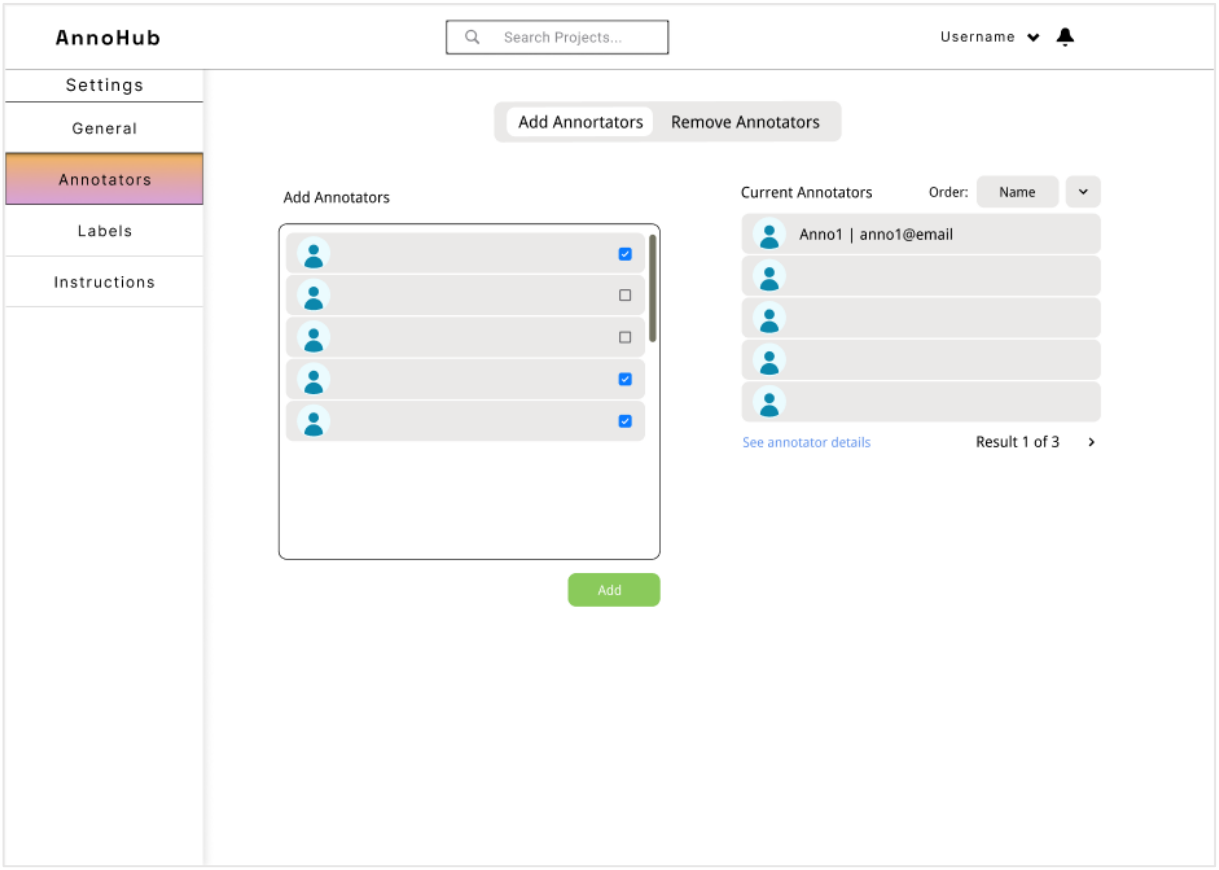
or

Drag and drop files here

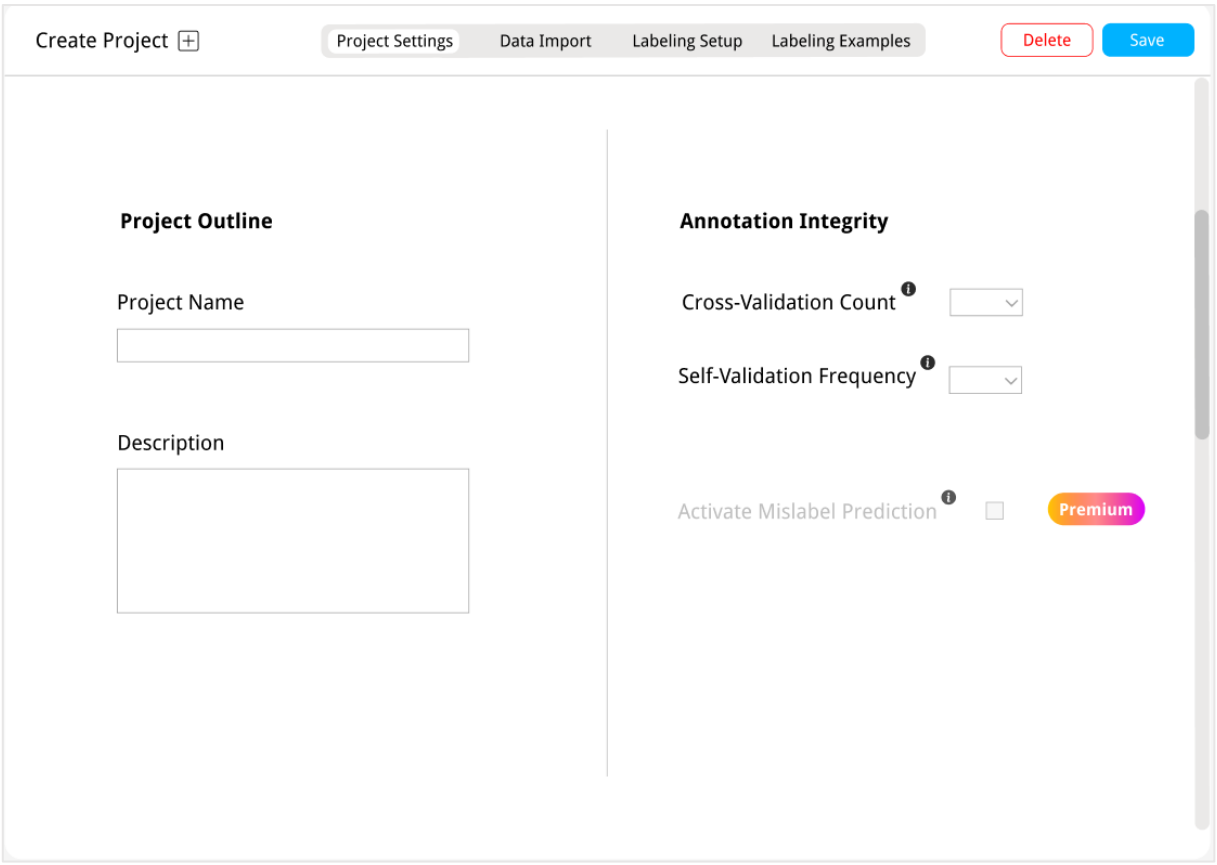
3.5.5.2 Publisher Project Settings

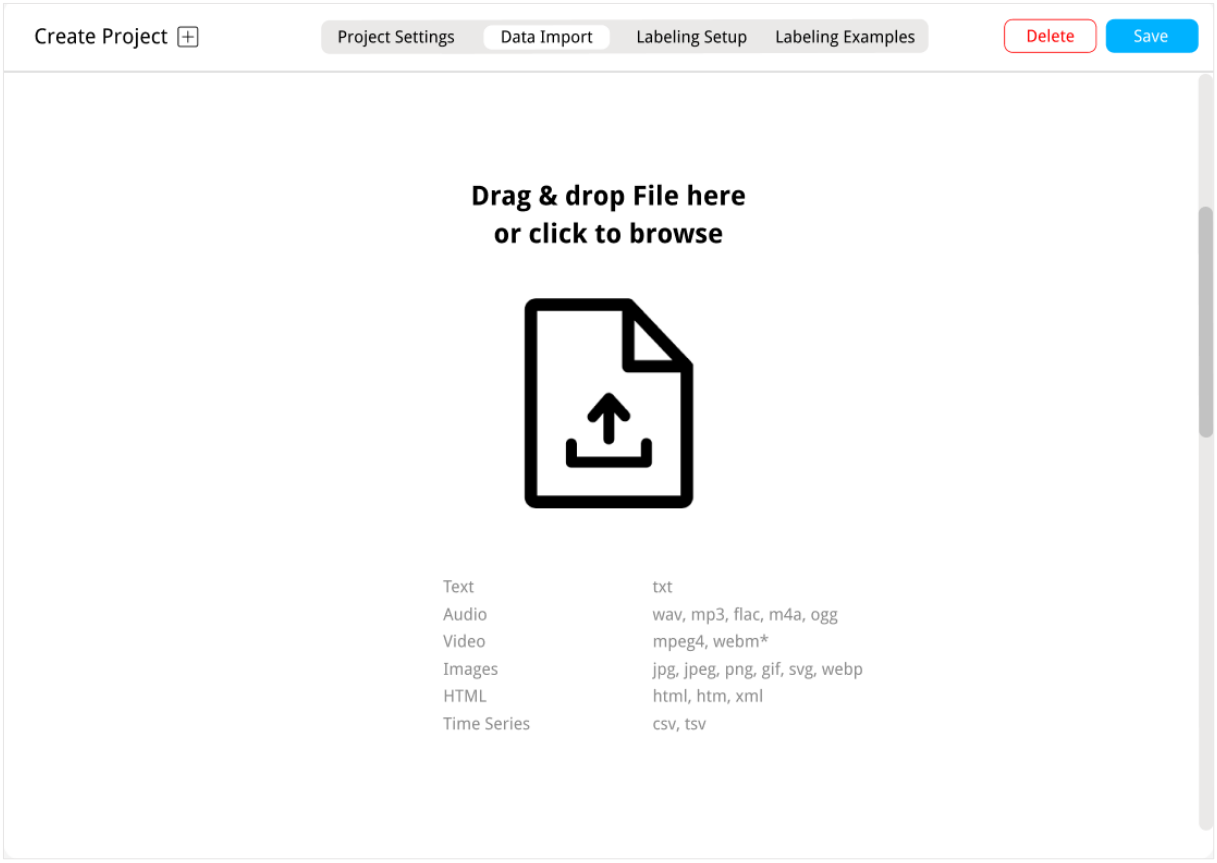


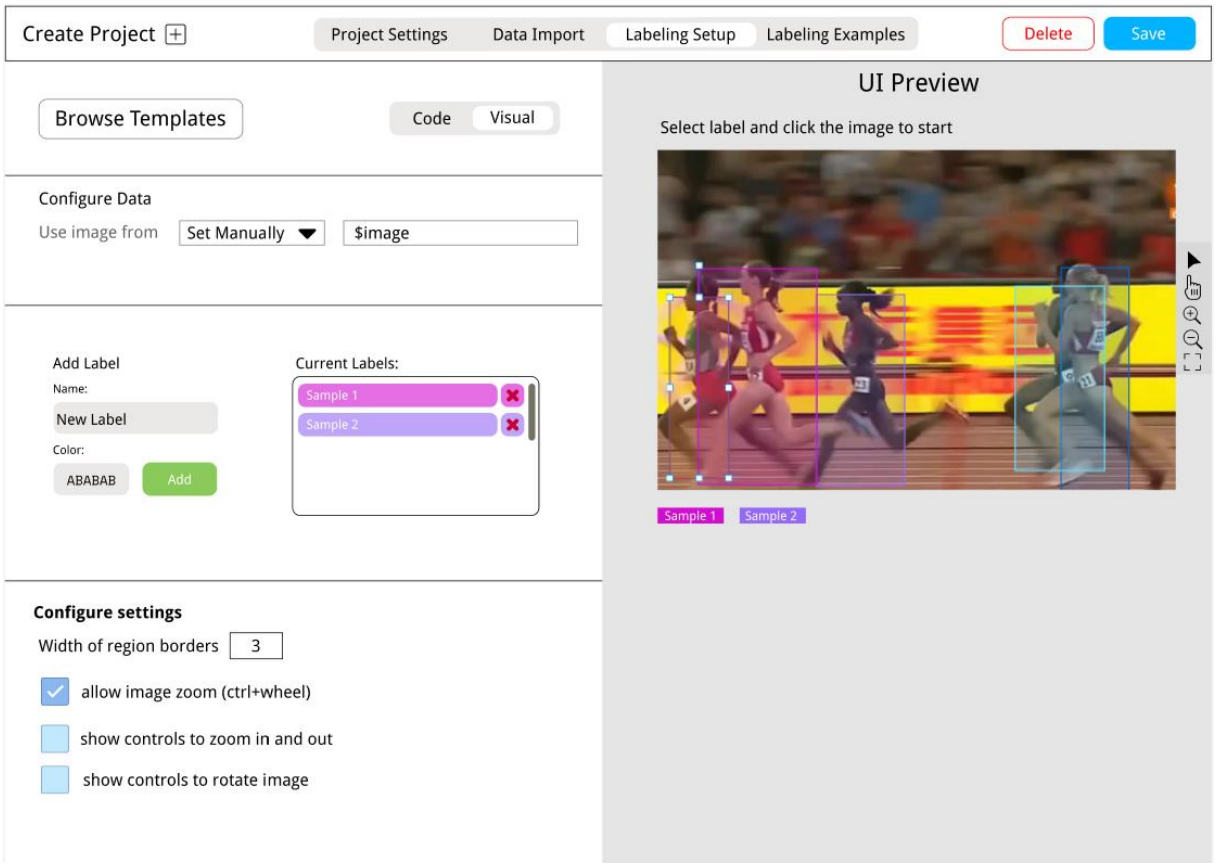



AnnoHub
Search Projects...
Username
Settings
General
Annotators
Labels
Instructions
Add Annortators
Remove Annotators
Add Annotators
Add
Current Annotators
Order:
Name
Anno1 | anno1@email
See annotator details
Result 1 of 3

3.5.5.3 Publisher Create Project




Create Project
Project Settings
Data Import
Labeling Setup
Labeling Examples
Delete
Save
Drag & drop File here
or click to browse
Text txt
Audio wav, mp3, flac, m4a, ogg
Video mpeg4, webm*
Images jpg, jpeg, png, gif, svg, webp
HTML html, htm, xml
Time Series csv, tsv


Create Project
Project Settings
Data Import
Labeling Setup
Labeling Examples
Delete
Save
Browse Templates
Code
Visual
Configure Data
Use image from
Set Manually
$image
Add Label
Name:
New Label
Color:
ABABAB
Add
Current Labels:
Sample 1
Sample 2
Configure settings
Width of region borders
3
allow image zoom (ctrl+wheel)
show controls to zoom in and out
show controls to rotate image
UI Preview
Select label and click the image to start
Sample 1
Sample 2

Create Project

Project Settings | Data Import | Labeling Setup | Labeling Examples

Delete | Save

- Computer Vision
- NLP
- Audio/Speech
- Conversational AI
- Ranking & Scoring
- Structured Data P.
- Time Series Analys.
- Videos
- Generative AI

Custom Template

Labeling Instructions

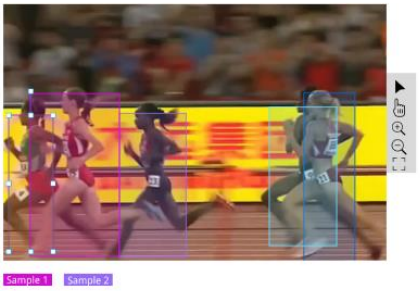
Instructions

1. Add a new label.
   a. Look for the option to add a new label.
   b. Click on Add Label button
2. Label selection
   a. Once the label is added, click on the object that corresponds to the label.
   b. If applicable, draw a bounding box around the object.
3. Multiple labels
   a. If there are multiple objects that need to be labeled, repeat steps 2 and 3 for each object
4. Review and double-check
   a. Before finalizing, review the entire image to check for any missed objects or labeling mistakes.

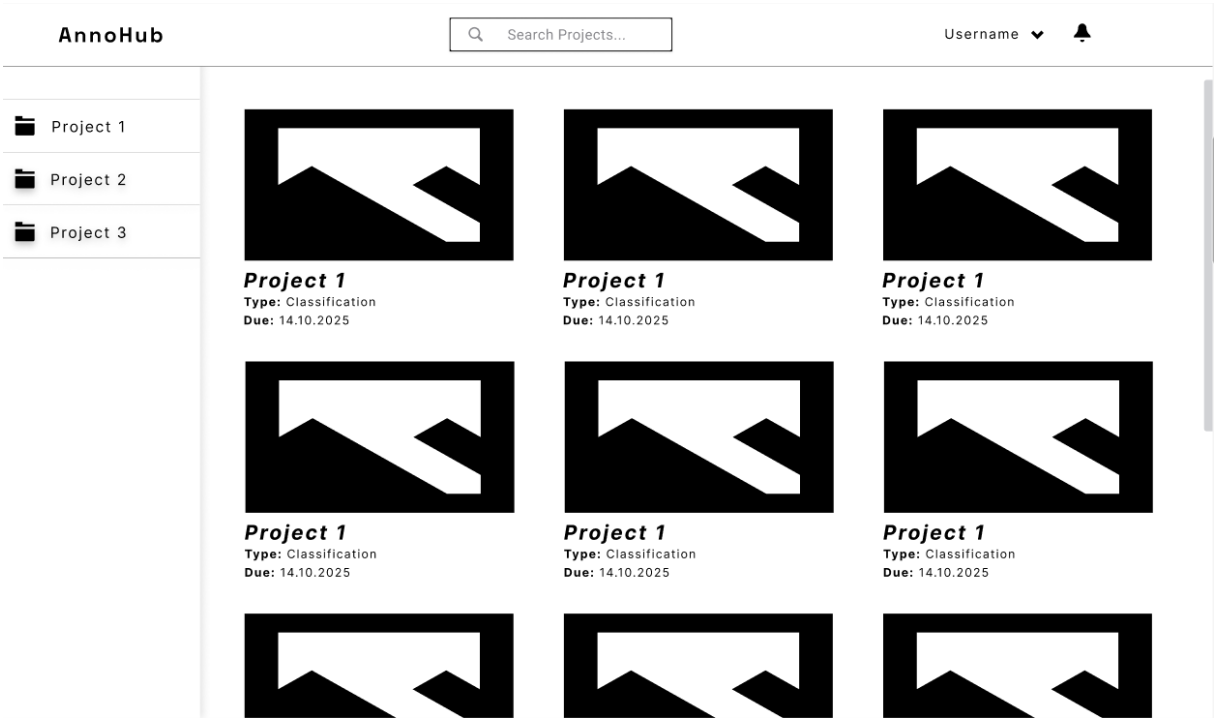
You may practice by labeling below image.

Example Labeling

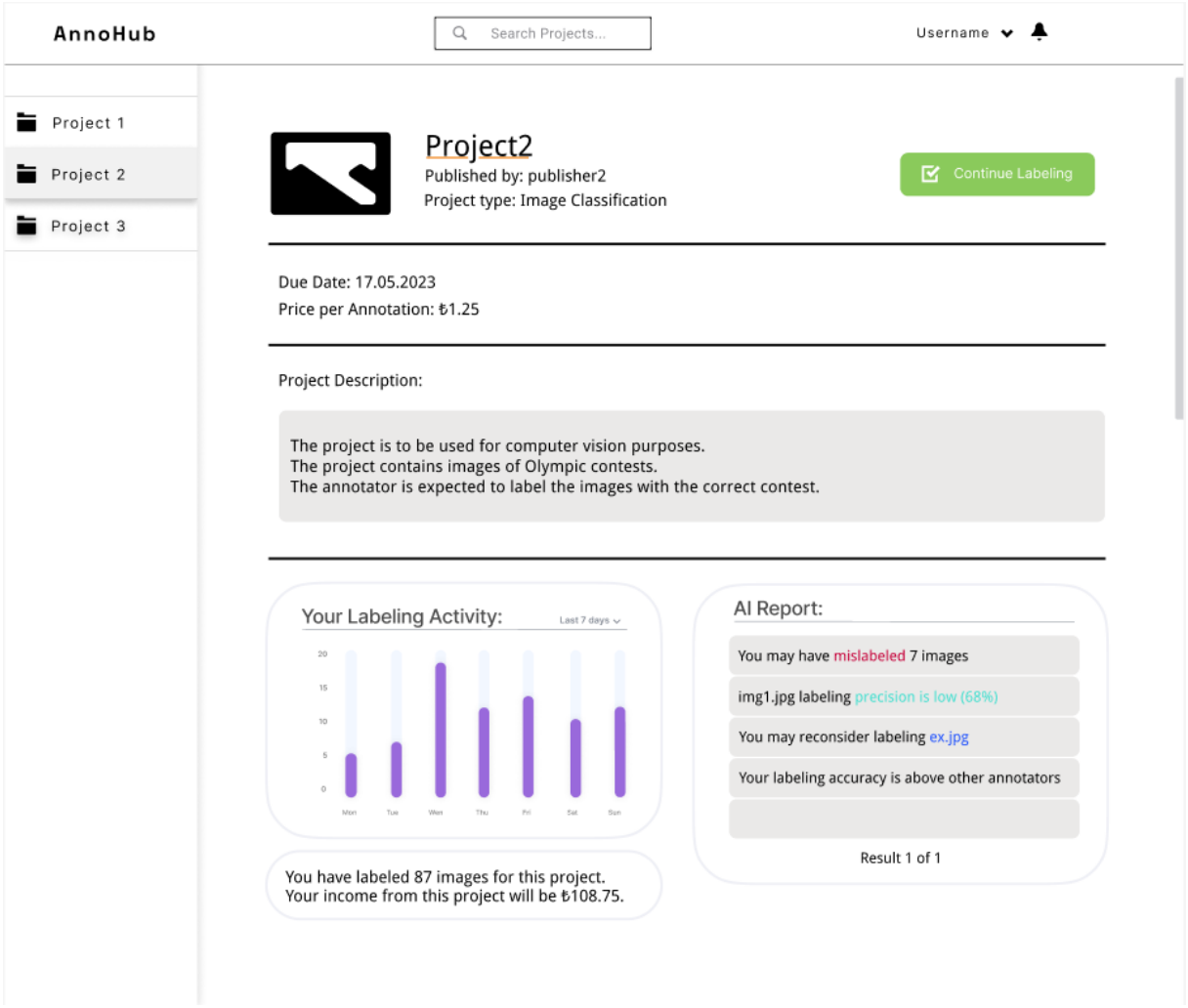
Here, you should explain the annotators clearly how and why they should label the data with respect to the instructions you gave above.



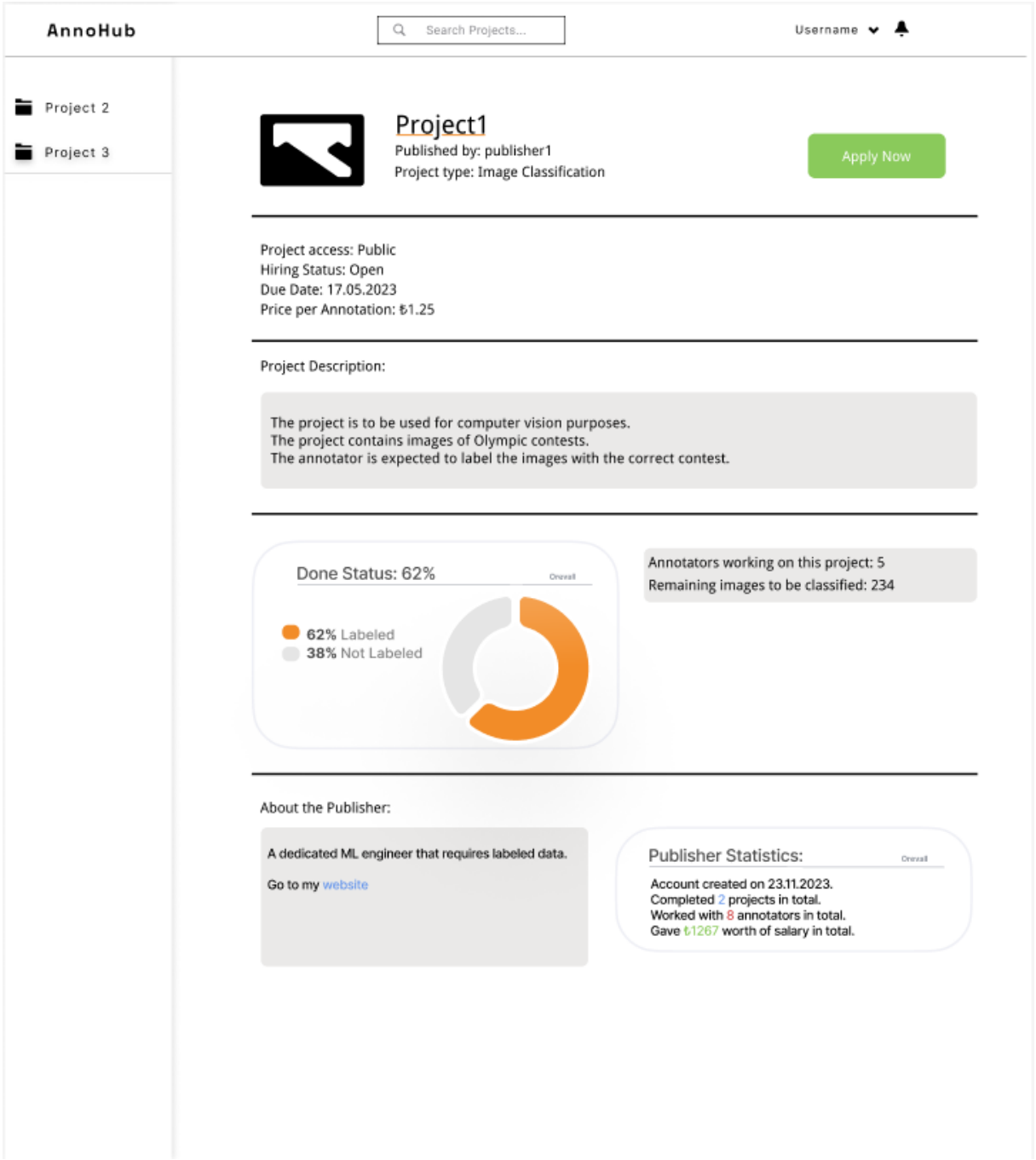
3.5.5.4 Labeler Dashboard



3.5.5.5 Labeler Project Page



3.5.5.6 Labeler Apply Project



TERMS OF USE

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis laoreet, enim sed tempus sodales, turpis massa dignissim nunc, a feugiat lacus velit ac turpis. In sed lacus sed tellus facilisis semper vitae ut nulla. Praesent tempor aliquam odio ut mattis. Integer at metus et sapien congue vestibulum ut sed ligula. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Nam feugiat rhoncus ligula ac mollis. In ex est, rhoncus nec metus quis, tempor posuere lorem. Vivamus placerat vitae ex sed cursus.

Quisque efficitur felis quam. Etiam vulputate sit amet tortor sed porta. Quisque ultrices, arcu sit amet faucibus rutrum, justo nisl volutpat lorem, sed maximus ligula sem mattis erat. Nam non vulputate lacus. Nunc fermentum rutrum orci, et lacinia neque luctus at. Phasellus feugiat vitae lacus consequat condimentum. Suspendisse dapibus fringilla pharetra. Fusce sit amet odio neque. Cras at felis dictum, ultrices urna nec, tristique nisi. Donec condimentum scelerisque lorem. Mauris in urna ac tortor scelerisque mattis et ut erat. Nam non neque velit. Sed non justo ut felis tincidunt mollis nec vitae lectus. Quisque in lectus ac purus elementum cursus. Aenean sed nulla nisl. Curabitur congue eleifend risus sit amet iaculis.

Duis vulputate lorem sed sapien mollis, eu mattis magna venenatis. Integer tempor libero fringilla mauris blandit posuere. Nam convallis tempus nulla, in tincidunt augue vehicula vitae. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Nam vitae facilisis massa. Donec laoreet lorem vitae est dignissim commodo. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia curae; Etiam ut tincidunt sapien, id luctus enim.

Cras non lorem quis sem porta porttitor. Pellentesque dictum sagittis bibendum. Cras et vestibulum nisi, et consectetur massa. Aenean ut dui consequat, sagittis eros vel, laoreet leo. Aenean non arcu nibh. Morbi vel ante ac tortor cursus ullamcorper. Donec tempus dignissim rutrum. Maecenas commodo massa libero, ac suscipit ex congue in. Etiam ut mauris tincidunt, finibus nisl ut, vehicula dui. In interdum porta sapien vitae semper. Cras a tempus felis.

Nulla quam nibh, pulvinar non efficitur vel, varius at felis. Duis vehicula justo ut commodo congue. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia curae; Nam eu rhoncus diam. Maecenas tempor efficitur sodales. Donec porta lectus ut dictum cursus. Proin rhoncus lectus ut libero condimentum, vel ultricies libero blandit. Nunc ultrices lacinia arcu, quis sagittis lorem consequat vitae. Vivamus semper sollicitudin ligula, in pharetra nibh aliquet id. Praesent id facilisis ex, non aliquam diam. Vivamus eleifend, est eu lobortis auctor, tortor nisl faucibus dui, nec pellentesque felis turpis condimentum nisi. Ut ipsum nulla, cursus a elit ac, dictum congue sem.

I understand and accept the terms of use

LABELING TEST

1. Please label each runner in the below image.

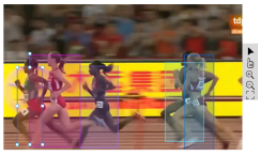
Add Label Names

Add

Labels (2)

Sample 1

Sample 2



2. Please label the words that indicate the time/date in the below paragraph.

On a balmy summer afternoon, we embarked on a spontaneous adventure to the beach. The day, marked by the vibrant hues of the azure sky and the gentle warmth of the sun, seemed to unfold like a story. Glancing at the date on our phones, we realized it was the first day of August. With the date in mind, we spread our beach towels and reveled in the simple joys of sand between our toes and the rhythmic sounds of waves kissing the shore. As the sun dipped lower on the horizon, casting a warm orange glow, we couldn't help but appreciate the beauty of the moment and the fleeting nature of time. The date on the calendar acted as a subtle reminder of the temporal nature of our beachside escapade, making it all the more precious. And so, as the day waned and the stars began to emerge, we gathered our belongings, carrying with us the memories of a sun-soaked summer day that unfolded with the grace of an idyllic chapter in time.

Upload Supporting Files

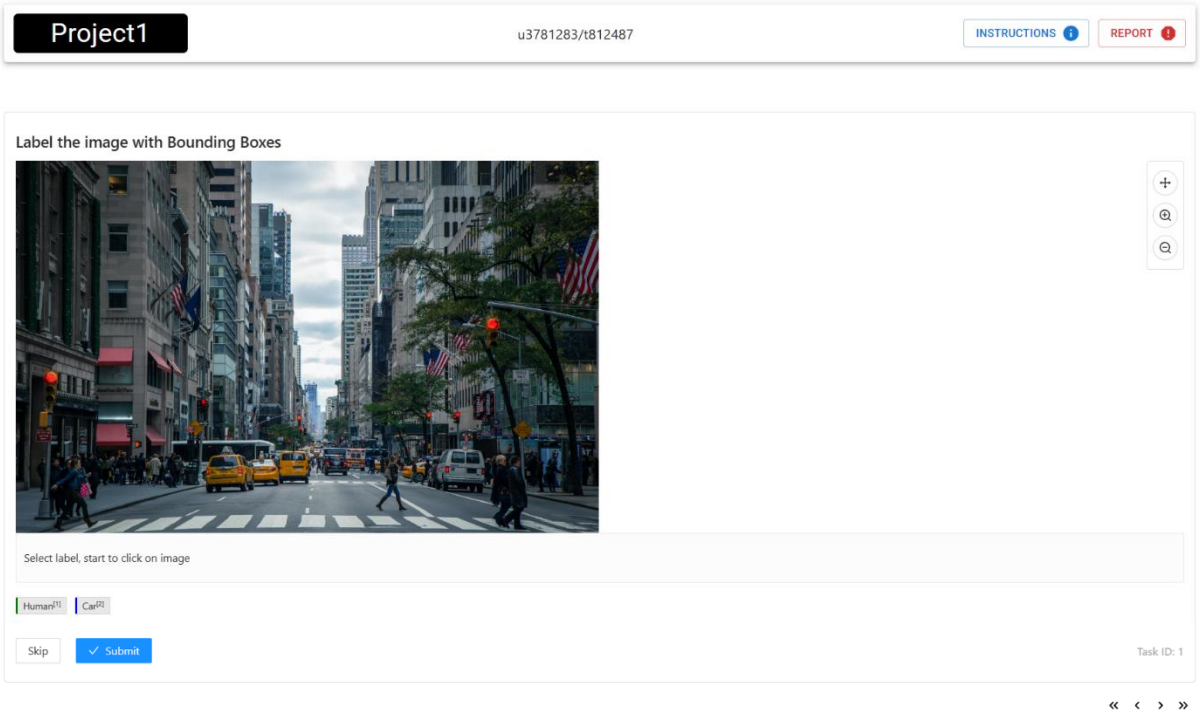
Choose files to upload

Browse Files

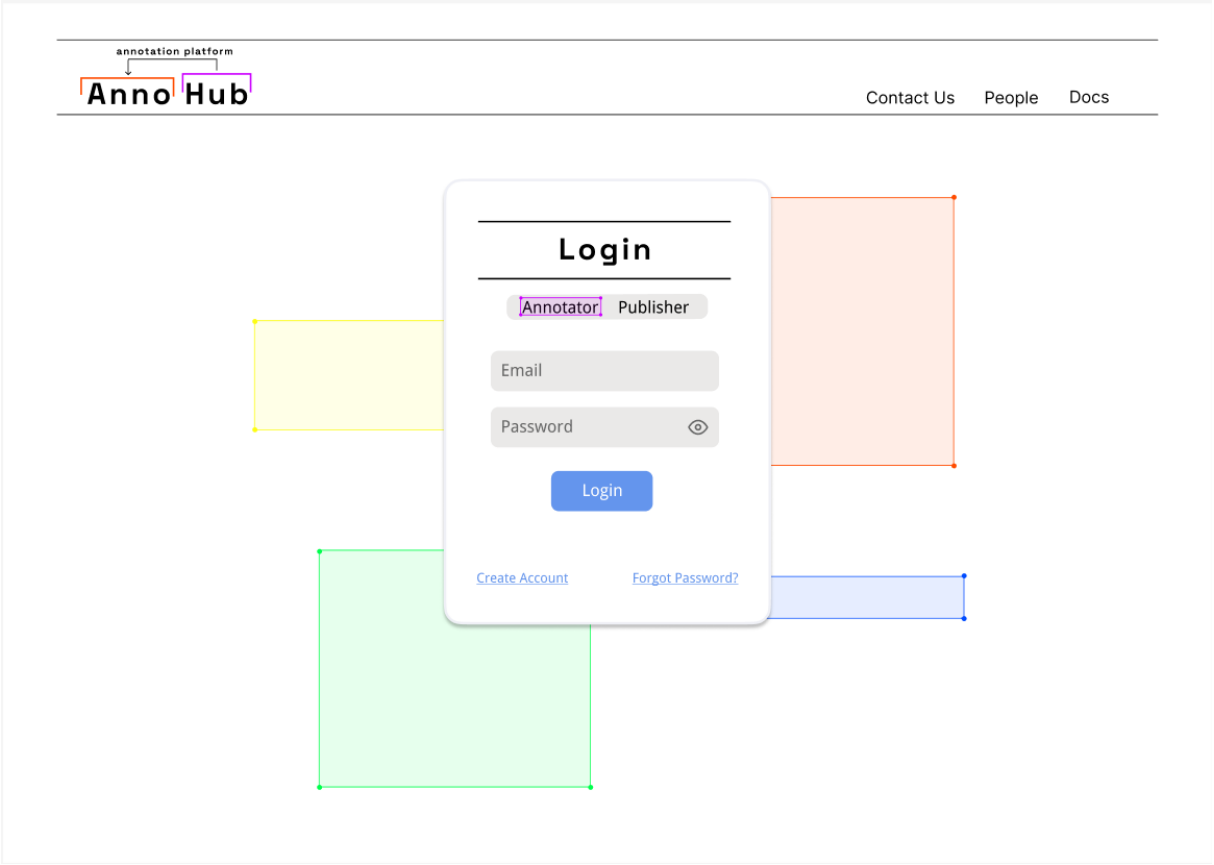
or

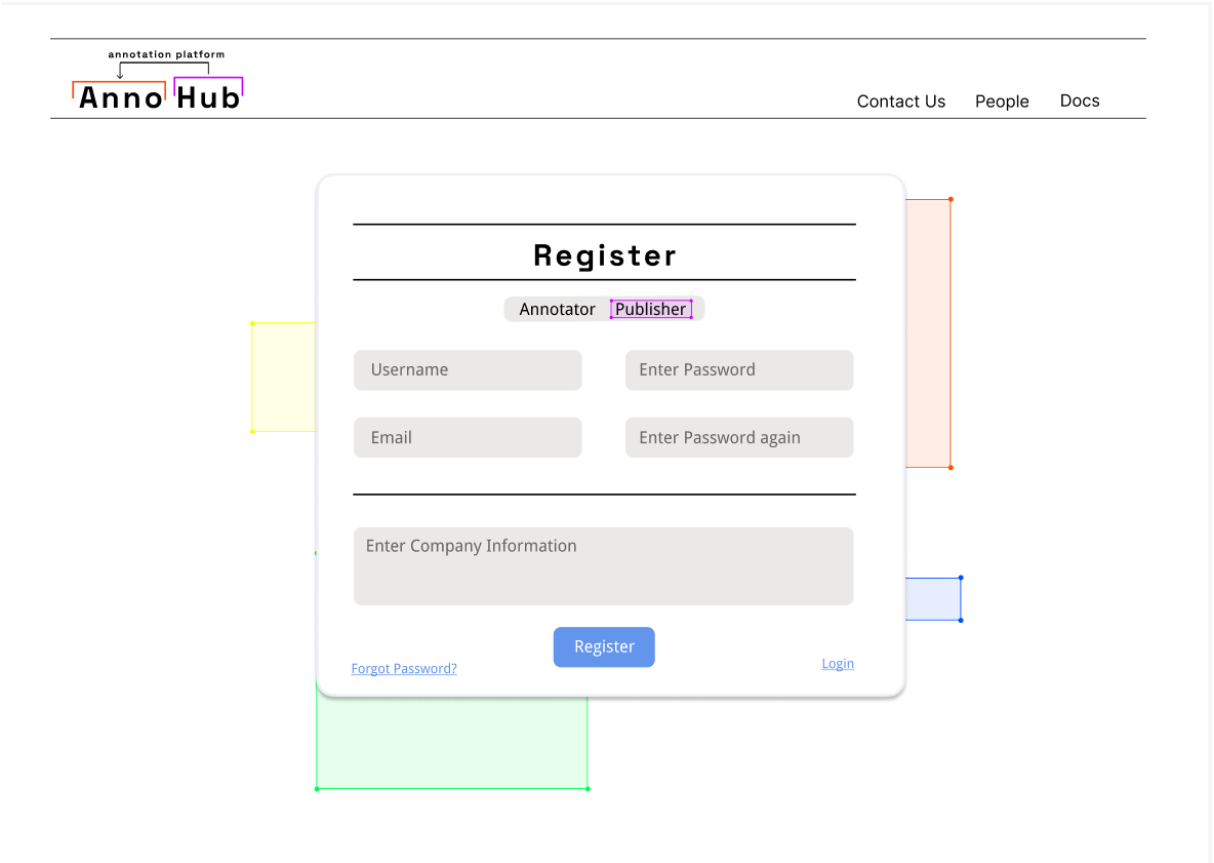
Drag and drop files here

3.5.5.7 Labeler Label Data



3.5.5.8 Login / Register





4. Other Analysis Elements

4.1 Consideration of Various Factors in Engineering Design

4.1.1 Constraints

4.1.1.1 Operational & Implementation Constraints

- AnnoHub will be a web application.
- GitHub will be used for team collaboration, code implementation, and version control.
- Python Flask will be used for the backend implementation and React will be used for the frontend implementation. This was found suitable for agile development.
- MongoDB will be used for data annotation because a NoSQL database allows flexibility in handling frequently changing schemas. PostgreSQL will be used to store user information since it performs better when dealing with bigger datasets.
- A cloud service (such as Amazon Web Services) is planned to be used for hosting.
- Electron JS will be used for the on-premise desktop application implementation because it is interpreted and its HTML-CSS-JS implementation is beneficial for agile development.
- Customers are expected to have decent servers to ensure data privacy while keeping the application functional.
- There should exist customers and labelers for AnnoHub's operational purposes.
- The data to be labeled does not have to be hosted in AnnoHub. However; in the event of customers choosing to keep their data private, certain features -such as mislabel prediction- are either going to be disabled or the source code will be shared with the customer through the on-premise application since both cannot co-exists.

4.1.1.2 Economic Constraints

- The main economic constraint for AnnoHub is the expenses that are associated with the server maintenance. This includes costs related to hosting, cloud services, and storage. Server resources must be efficiently managed to control these expenses and ensure that AnnoHub is functioning smoothly. The storage costs are expected to be higher than the others and should be given importance.
- Another economic constraint is the costs related to financial transactions, like money transfers to labelers. These transactions involve fees and currency conversion charges, which need to be considered in the financial planning of AnnoHub.
- Initial investment is another significant economic constraint. While the project can operate with a moderate initial investment, future financial planning is essential. Adequate funding is needed not only for the initial setup costs but also for the ad expenses and other ongoing needs.

4.1.1.3 Legal & Ethical Constraints

- A mandatory objective of AnnoHub is to stay observant to the Turkish Personal Data Protection Law (KVKK).
- User-provided information must not be leaked to any third-party organization unless explicitly approved by the user and is within KVKK's scope.
- The user's stored data should be destroyed within a time interval or when the user requests the data's destruction. The detailed policies on data storage and destruction must be provided to the user as stated in KVKK.
- For the data to be stored and processed, several clarification and express consent texts must be given to the user. The methodologies used to store/process user's data should be introduced to the user together with the other are legal and ethical considerations. The user must approve these texts with his own consent before using the application.
- The data must be stored/processed exactly the way that was introduced to the user which gets the user's approval.

4.1.2 Standards

4.1.2.1 Compliance with KVKK entails

- Registering with the Data Controllers' Registry (VERBIS) before processing any personal data.
- Obtaining explicit consent from data subjects for processing their personal data, except in specific circumstances outlined by the law.
- Ensuring the rights of data subjects are protected, including the right to be informed about data processing activities, the right to access their data, and the right to request correction or deletion of their data.
- Implementing measures for the secure transfer of personal data to foreign jurisdictions, which may involve ensuring adequate protection in the recipient country or obtaining authorization from the Turkish Data Protection Authority (DPA).
- Promptly notifying the DPA and affected data subjects in the event of a data breach.

4.1.2.2 Data Security and Privacy Standards

- ISO/IEC 27001: This international standard outlines the specifications for an information security management system (ISMS). Compliance with ISO/IEC 27001 will help to safeguard sensitive client data as it flows through our system to labelers.
- General Data Protection Regulation (GDPR): Adherence to GDPR is critical for protecting the personal information of individuals within the European Union. Our

system will implement necessary controls to manage user consent, data minimization, and individuals' rights over their data.
- Health Insurance Portability and Accountability Act (HIPAA): If dealing with health-related data, compliance with HIPAA will ensure the protection of personal health information, which could be part of the datasets being labeled.

4.1.2.3 Data Handling and Distribution Standards

- Data Minimization and Anonymization: Following best practices for data minimization and anonymization will ensure that labelers receive only the data necessary to perform their tasks, without exposing sensitive information.
- End-to-End Encryption: The use of end-to-end encryption protocols for data transmission will ensure that client data is secure in transit, preventing unauthorized access.

4.1.2.4 Software and Web Development Standards

- OWASP Top 10: We will adhere to the Open Web Application Security Project's Top 10 list to protect against common web application security risks.
- W3C Web Standards: Our user interface will align with the World Wide Web Consortium (W3C) guidelines, including HTML5, CSS3, and Web Content Accessibility Guidelines (WCAG), to ensure compatibility and accessibility.

4.1.2.5 API and Data Interchange Standards

- -RESTful API Security: Our APIs will follow REST security best practices, including authentication tokens, rate limiting, and proper session management.
- -JSON Web Tokens (JWT): For secure transmission of information between parties, JWT will be used, allowing us to safely pass user and session data.

4.1.2.6 Database Design and Management Standards

- SQL:2016 / NoSQL Best Practices: We will follow the latest SQL standards or NoSQL best practices to ensure efficiency, reliability, and integrity in data storage and retrieval.
- ACID Properties: Transaction handling within our databases will conform to ACID properties to ensure that all data transactions are processed reliably.

4.1.2.7 Compliance with Industry-Specific Standards

- Payment Card Industry Data Security Standard (PCI DSS): If our platform handles transactions, we will comply with PCI DSS to secure cardholder data.

4.2 Competitors

The utilization of AI and machine learning methodologies is expanding across diverse sectors, leading to an increased demand for high-quality labeled datasets. As of 2022, the global machine learning market attained a valuation of USD 36.73 billion, with an anticipated compound annual growth rate (CAGR) of 34.8% from 2023 to 2030 [1]. Consequently, it is plausible that existing tools in the market exhibit similar functionalities to our application. To address this, we conducted a comprehensive market analysis, identifying key competitors for AnnoHub. Following the research, three prominent tools - Labelbox, Label Studio, and Amazon Sagemaker Ground Truth- emerged as primary contenders due to their widespread usage and analogous functionalities.

Labelbox, LabelStudio, and Amazon Sagemaker Ground Truth are all comprehensive data labeling tools, they come with their advantages and disadvantages. Labelbox and Amazon Sagemaker are known for their prices being on the higher end after passing a limit, and it is difficult to pay for some publishers. Moreover, Labelbox and Amazon Sagemaker partially support outsourcing labeling, neither provides a comprehensive public workforce system. Although Amazon employs a similar system known as Amazon Mechanical Turk, our project's innovation expert, Ahmet Kocamaz, expressed dissatisfaction with the service due to inadequate information about labeled data metrics. Mechanical Turk is also notorious for not returning to the applications of the public. To solve these issues, AnnoHub aims to establish a public workforce, allowing customers the option to label their data through this workforce. Customers can select qualified public labelers via proficiency tests tailored to the labeling task, and publishers also can require specific documents from the labelers. AnnoHub strives to offer both cost-effective labeling and job opportunities through this system, a distinctive feature absent in competitors. By this system, AnnoHub will have a massive advantage over the use in academia and startup companies, because of the cheaper pricing enabled by the public workforce.

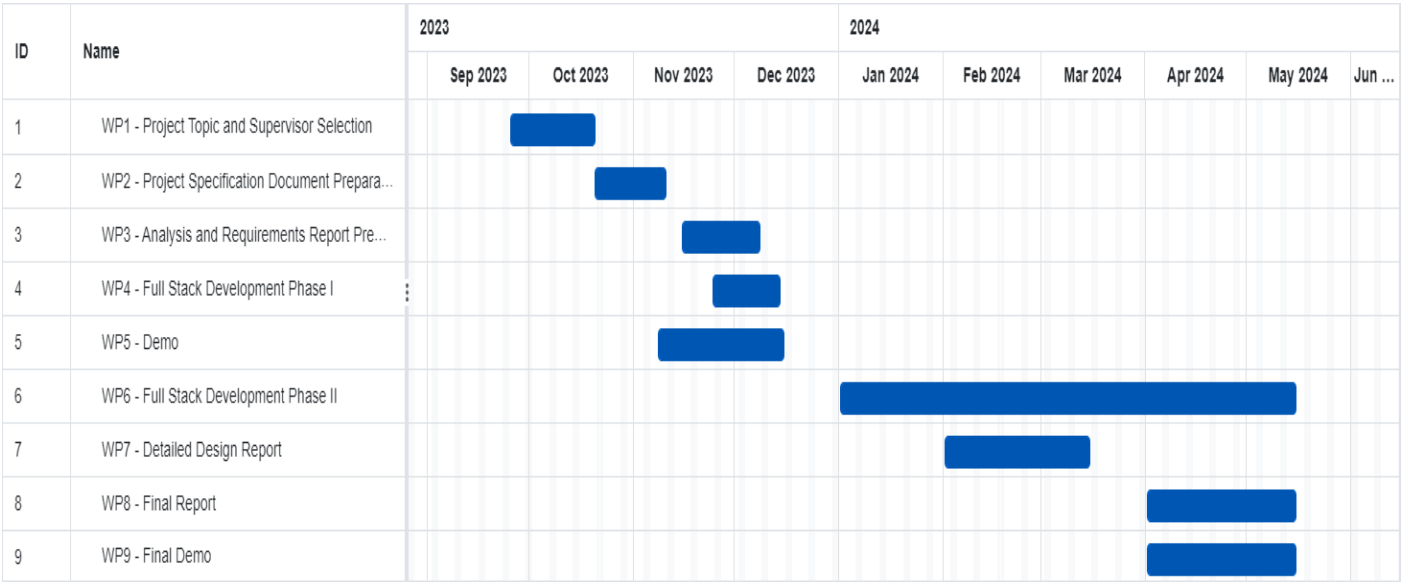
Another primary challenge addressed by our application is the facilitation of data labeling without the necessity of sharing sensitive information with external servers. Labelbox mandates customers to upload data on Google Cloud Services, storing it compulsorily in the US [2]. Catering to clients unwilling or unable to share data beyond their national borders (e.g., due to KVKK laws in Türkiye), AnnoHub offers a solution by maintaining only references to sensitive data on its servers and sharing the actual data

exclusively with designated labeler employees. Consequently, AnnoHub ensures privacy by design, alleviating the need to entrust external entities with sensitive information.

AnnoHub will not support model training like Amazon Sagemaker Ground Truth, but it will match the labeling functionalities by providing interfaces for various data types. Moreover, AnnoHub employs machine learning algorithms, including cross-validation, self-validation, and fraud detection algorithms. With the help of these algorithms, AnnoHub will also provide extensive analytics and metrics to the publishers. These mechanisms, coupled with the public workforce system, underscore AnnoHub's commitment to ensuring efficient and effective labeling, ultimately contributing to heightened customer satisfaction.

4.3 Project Plan

4.3.1 Gantt Chart



4.3.2 Work Packages

WP#	Work Package	Leader
WP1	Project Topic and Supervisor Selection	Emirkan Derköken
WP2	Project Specification Document	Onurcan Ataç
WP3	Analysis and Requirements Report	Dağhan Ünal
WP4	Full Stack Development Phase I	Emirkan Derköken
WP5	Demo	İlker Özgen
WP6	Full Stack Development Phase II	Bora Yılmaz
WP7	Detailed Design Report	Onurcan Ataç
WP8	Final Report	Bora Yılmaz
WP9	Final Demo	Dağhan Ünal

WP1: Project Topic and Supervisor Selection	
Start Date: 25 September 2023	**End Date:** 20 October 2023
Leader: Emirkan Derköken	**Members Involved:** All Members
Objectives: Finalizing the project topic and the project supervisor.	
Tasks: **Task 1.1 Project Topic Selection:** Brainstorm potential project topics. Conduct market and feasibility research about the topics that have been found. Select one of the topics that aligns the best. **Task 1.2 Project Supervisor Meeting:** Schedule and conduct a meeting with the supervisor. Present the project topic and get feedback. **Task 1.3 Innovation Expert Meeting:** Arrange a meeting with the innovation expert to get ideas and feedback related to the chosen project topic.	
Deliverables: **D1.1:** Project Information Form **D1.2:** Assessment of Innovation Form	

WP2: Project Specification Document Preparation	
Start Date: 20 October 2023	**End Date:** 10 November 2023
Leader: Onurcan Ataç	**Members Involved:** All Members
Objectives: Completing and delivering the Project Specification Document.	
Tasks: **Task 2.1 Writing Introduction:** Write an introduction for the Project Specification Document. Clearly outline the purpose, scope, and objectives of the project. **Task 2.2 Writing Design Requirements:** Specify the design requirements of the project. Address any design constraints in the project's development. **Task 2.3 Writing Feasibility Discussion:** Write a detailed analysis of the project's feasibility, considering technical, economic, legal, and operational aspects.	
Deliverables: **D2.1:** Project Specification Document	

WP3: Analysis and Requirements Report Preparation	
Start Date: 15 November 2023	**End Date:** 1 December 2023
Leader: Dağhan Ünal	**Members Involved:** All Members
Objectives: Completing and delivering the Analysis and Requirements Report.	
Tasks: **Task 3.1 Writing Introduction and Problem Definition:** Introduce the Analysis and Requirements Report by outlining the project's background and clearly defining the problem. **Task 3.2 Use Case Diagram:** Develop a detailed use case diagram to illustrate system functionality from the user's perspective. **Task 3.3 Activity Diagrams:** Create activity diagrams to represent the dynamic aspects of the system, showing the flow of activities and actions in the system. **Task 3.4 State Diagrams:** Create state diagrams to model the different states in the system and the transitions between these states. **Task 3.5 API Design:** Define the API for the project, specifying the methods and protocols for communication between different software components. **Task 3.6 Database Design Diagram:** Create an ER diagram outlining the structure and relationships between different data entities. Address database normalization. **Task 3.7 Designing The User Interface:** Create a user-friendly interface design. Consider user experience principles and aesthetics of the interface. **Task 3.8 Project Plan:** Make a detailed project plan covering tasks, milestones, timelines, and team member allocation.	

Deliverables: **D3.1:** Analysis and Requirements Report

WP4: Full Stack Development Phase I	
Start Date: 25 November 2023	**End Date:** 14 December 2023
Leader: Emirkan Derköken	**Members Involved:** All Members
Objectives: Completing the MVP of the application with basic features implemented.	
Tasks: **Task 4.1 Front-end Development:** Begin the development of the user interface and components. Implement responsive design and navigation. **Task 4.2 Back-end Development:** Begin the development of server-side functionalities, databases, and application logic. **Task 4.3 Database Setup:** Establish the database, including schema creation.	
Deliverables:	

WP5: Demo	
Start Date: 8 November 2023	**End Date:** 15 December 2023
Leader: İlker Özgen	**Members Involved:** All Members
Objectives: Preparing the demo and the presentation.	
Tasks: **Task 5.1 Decision on the Demo Sequence:** Decide on the sequence of what will be shown for the demo, considering the logical flow of key features. **Task 5.2 Presentation Preparation:** Prepare a presentation for the demo, focusing on key project features and user interactions. **Task 5.3 Rehearsals:** Conduct rehearsals for the demo presentation to practice timing and team coordination.	
Deliverables: **D5.1:** Project Presentation **D5.2:** Project Demo	

WP6: Full Stack Development Phase II	
Start Date: January 2024	**End Date:** Mid-May 2024
Leader: Bora Yılmaz	**Members Involved:** All Members
Objectives: Complete the implementation of the application, and prepare the release version.	
Tasks: **Task 6.1 Continuing Development Activities:** Continue the full-stack development to implement additional features and functionalities. **Task 6.2 Integration Testing:** Do testing to ensure the integration between the front-end and back-end components is successful. **Task 6.3 Performance Optimization:** Optimize the performance of the system, focusing on responsiveness, speed, and resource allocation. **Task 6.4 Security Implementation:** Integrate security measures into the application, including data encryption, user authentication, and authorization mechanisms.	
Deliverables: **D6.1:** The final version of the application	

WP7: Detailed Design Report	
Start Date: February 2024	**End Date:** Mid-March 2024
Leader: Onurcan Ataç	**Members Involved:** All Members
Objectives: Completing and delivering the Detailed Design Report.	
Tasks: **Task 7.1 Writing Design Goals:** Clearly explain the design goals of the project, outlining the desired outcomes and functionalities. **Task 7.2 Proposed Software Architecture:** Present a detailed overview of the proposed software architecture, highlighting the key features. **Task 7.3 Deciding on Test Cases and Write Tests:** Create test cases to validate the functionality, performance, and security aspects of the software. **Task 7.4 Writing Constraints and Standards:** Identify and highlight the constraints and standards that effect the software design and development process. Ensure compliance with the standards and legal requirements.	
Deliverables: **D7.1:** Detailed Design Report	

WP8: Final Report	
Start Date: April 2024	**End Date:** Mid-May 2024
Leader: Bora Yılmaz	**Members Involved:** All Members
Objectives: Completing and delivering the Detailed Final Report.	
Tasks: **Task 8.1 Final Architecture and Design Details:** Provide an overview of the final software architecture and design in detail. **Task 8.2 Development/Implementation Details:** Highlight the implementation details of the project, including coding standards and version control practices. **Task 8.3 Maintenance Plan and Details:** Outline a detailed plan for the maintenance of the software. Specify procedures for bug fixes and updates. Consider scalability in the maintenance plan. **Task 8.4 Conclusion and Future Work:** Conclude the final report by summarizing key achievements and the overall project outcome. Discuss potential future enhancements related to the project.	
Deliverables: **D8.1:** Final Report	

WP9: Final Demo	
Start Date: April 2024	**End Date:** Mid-May 2024
Leader: Dağhan Ünal	**Members Involved:** All Members
Objectives: Completing and delivering the Final Report.	
Tasks: **Task 9.1 Decision on the Demo Sequence:** Decide on the sequence of what will be shown for the demo, considering the logical flow of key features. **Task 9.2 Presentation Preparation:** Prepare a presentation for the demo, focusing on key project features and user interactions. **Task 9.3 Rehearsals:** Conduct rehearsals for the demo presentation to practice timing and team coordination.	
Deliverables: **D9.1:** Project Presentation **D9.1:** Project Demo	

4.4 Ensuring Proper Team Work

Maintaining effective teamwork is one of our main concerns in the project's progress, for which we follow the principles below.

- Every member should attend the regular meetings.
- Every member should contribute to the decision-making process.
- Every member should take part in each step in the project's progress.
- Every member is responsible for doing the necessary research about their work.
- Every member should meet the deadlines that are determined in the regular meetings.
- The distribution of work in each step in the project's progress should be fair.

5. References

[1] “Machine learning market size, share & growth report, 2030,” Machine Learning Market Size, Share & Growth Report, 2030, https://www.grandviewresearch.com/industry-analysis/machine-learning-market#:~:text=Report%20Overview,how%20businesses%20and%20people%20operate (accessed Nov. 16, 2023).

[2] “Access, storage, and security,” Labelbox docs, https://docs.labelbox.com/docs/access-storage (accessed Nov. 16, 2023).