1 Introduction

The human epidermal growth factor receptor 2 (HER2) protein plays a critical role in the diagnosis and treatment of aggressive breast cancers. Over-expression of the HER2 protein occurs in around 20-30% of breast cancer tissues, and is associated with higher disease aggression, recurrence rate, and mortality [1]. In addition, HER2 over-expression is the only clinically usable biomarker for selection of targeted therapy in metastatic gastroesophageal adenocarcinoma [2]. Although HER2 amplification is less common in lung cancer, HER2-mutant lung cancer responds to HER2 inhibitors, and the response rate of such cancers to HER2 antibody drug conjugates is an active area of research [2]. Furthermore, targeted HER2 treatments have shown promising response rates in certain colorectal cancers [2]. In HER2 positive breast cancer, targeted therapies show significantly positive results in response rate, time to progression, survival advantage, and risk of recurrence [3]. Thus, the diagnosis of HER2 positive tumors is critical for selecting targeted treatments and ensuring optimal patient outcomes.

One technique for assessing a sample's level of HER2 over-expression is through immunohistochemistry (IHC) staining. In this technique, a sample is stained with chemically labeled antibodies in order to highlight the presence of antigens in the infected tissue [4]. A tissue's HER2 over-expression status is categorized on a scale of 0 to 3, with a score of 0 or 1 considered negative, 2 considered equivocal and requiring further testing, and 3 considered positive [3].

Currently, pathologists are trained to interpret HER2 through textbooks and labeled samples, as well as sessions with a senior pathologist. In these sessions, the trainee pathologist receives stained HER2 slides, interprets them, and then compares and discusses their results with the senior. We propose an online platform through which pathologists can rapidly assess a wide range of samples and review their assessments for accuracy. This platform provides an online, automated analog that could supplement or replace the traditional training process. Through this methodology, pathologists can improve their skills easily and remotely, and access a wide range of samples to evaluate. Engleberd et al. found a significant decrease in diagnostic error after subject participation in a similar execrise [5], so there is reason to believe that such exercises produce real-world improvements in diagnostic accuracy. With future work, pathologists will be able to develop their diagnostic skills on criteria other than HER2 over-expression with the same level of ease and accessibility.

2 Results

BioGames is a web application that provides users with a series of IHC stained HER2 samples, allowing them to indicate their diagnosis of the sample's HER2 status between 0 and 3 (s_g), and then review the cores that they labeled correctly and incorrectly. The game also assigns a number of points from 0 to 3 for each diagnosed sample, calculated as $|s_g - s_t|$, where s_t is the true HER2 status of the core.

A "game" in the application's terminology is a set of "challenges" presented to the user, where each challenge corresponds to one HER2 tissue sample. Each challenge is scored according to the formula above, and then the game's total score is computed as the sum of the point values of each individual challenge. The total time taken to complete each game is also recorded, and factors into a player's ranking on the scoreboard displayed at the end of each game.

The game begins when the player enters their username on the main menu and clicks "Play," at which point the game and challenge entities are created in the game's database. Then, the player is presented with the first challenge. After a two second waiting period, the player is able to score the HER2 core, and the next one is displayed (Fig. 1). After the player scores the last core, they are presented with their results, along with a scoreboard displaying the top players (Fig. 3). Players are ranked by their best game, ordered first by total score (descending), and then by their total time (ascending). The player is also presented with the option to play again.

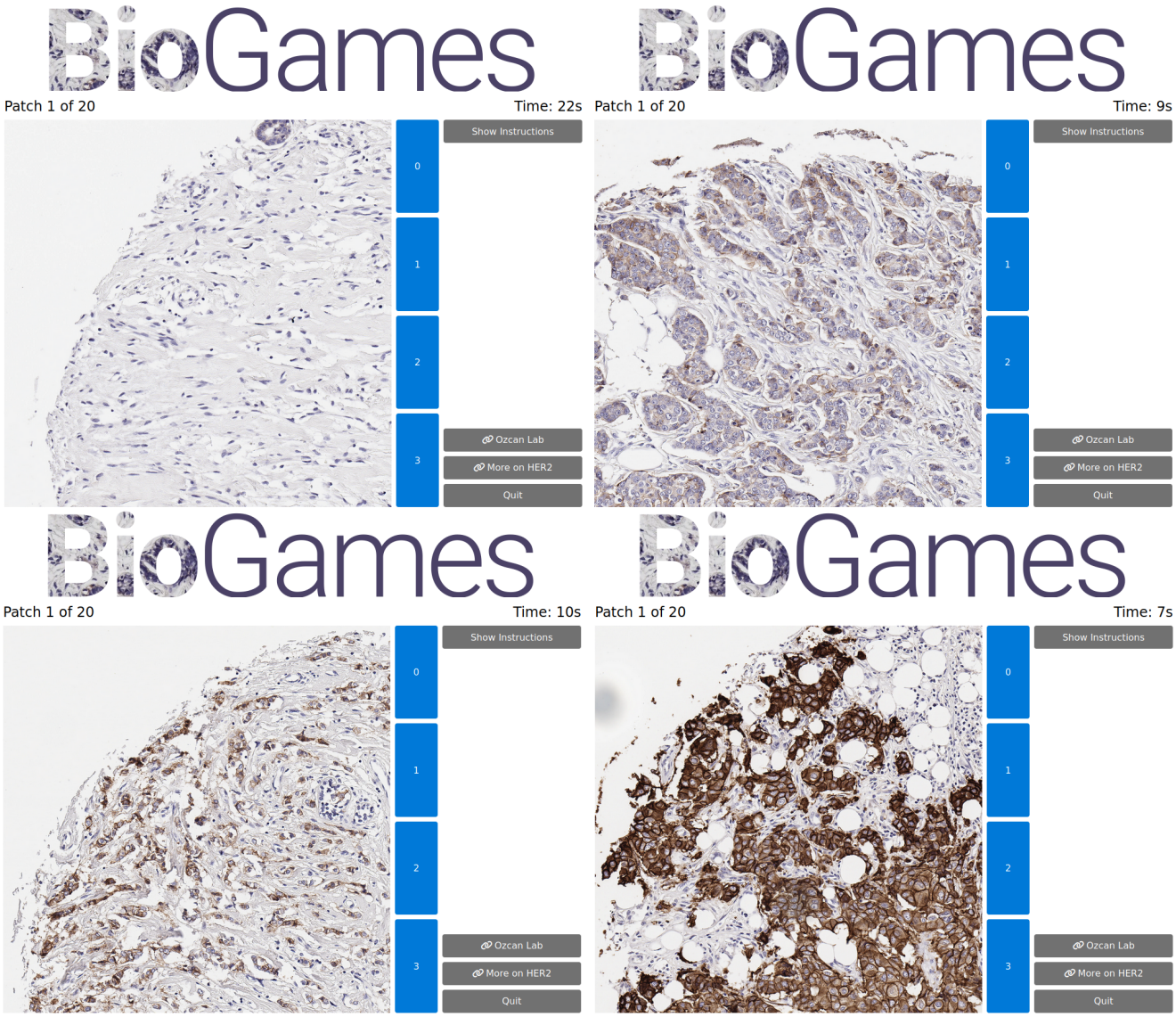

Figure 1: Four screenshots of the game, each with a patch with a different HER2 score (starting at zero in the top left screenshot and increasing left-to-right, then top-to-bottom

3 Methods

3.1 Data Procurement

We obtained fifteen unlabeled breast tissue microarray slides from TissueArray, each containing about 100-200 cores. These samples underwent HER2 immunhistochemical staining at the UCLA Translational Pathology Care Laboratory. We captured bright-field images of the stained slides usind the ZEISS AxioScan Z1 slide scanner with a x20/0.8NA Plan-Apo objective lens. The whole-slide images were then processed with a deep-learning based algorithm described in [6] to identify and extract each core.

3.2 Data Organization

While we obtained a large number of cores, we desired a higher number of images to present to the user for diagnosis. While no image will appear twice in the same game, having too few images would lead to a user memorizing images over multiple games, drastically reducing their identification time and artificially improving their position on the scoreboard. Taking this into consideration, we split each core into several patches. This presented the problem of patch diagnosability and disagreement between what HER2 score a patch might express and the core's actual score. For instance, a core may present a 3+ HER2 score in the top right quadrant, but 1+ in the lower left quadrant. A patch from the lower left quadrant would then differ in score from the core's labeled score (3+), so it became necessary to label each individual patch. To this effect, we used a deep-learning based classifier algorithm to label the patches, and included only patches that were identified as having a HER2 score congruent with their original cores. However, the classifier achieved less than 80% accuracy during validation, so greater care was needed to ensure an accurately labeled dataset. As such, each patch was then subjected to manual review by a trained pathologist, and discarded if the patch was undiagnosable or incongruent with its core. This process resulted in 10154, 3822, 7078, and 4190 patches with scores of 0, 1, 2, and 3, respectively. These patches were scaled down to 1024x1024 pixels and re-encoded from PNG to the WebP image format, reducing image size from roughly 20MB to roughly 250KB. This was done primarily to reduce the loading time for each image, but has the added benefit of reducing the amount of storage needed on the game's host machine.

3.3 Game Design

Code for the game consists of a single-page web application (SPA) written in TypeScript with the React UI library and a back-end application server written in Rust. Data is stored in a PostgreSQL database, and the HER2 core images are stored directly on the host machine's filesystem. The combination of TypeScript and React was chosen for their popularity and maturity. React is developed by Meta and was first released on July 2nd, 2013 [7], and TypeScript is developed by Microsoft and first appeared in 2012 [8]. PostgreSQL was also chosen for its maturity, originating in 1985 with the POSTGRES project at the University of California, Berkeley [9]. Like React, TypeScript, and Rust, PostgreSQL is an open source project, and all three are completely free to use. Rust is a general purpose programming language focused on performance and memory safety [10]. Memory safety is an increasingly important goal in software development, as the majority of security vulnerabilities are attributed to issues with memory safety. For example, roughly 70% of vulnerabilities at Microsoft and Google in recent years have been attributed to such issues [11]. Furthermore, the United States White House Office of the National Cyber Director has recently released a report recommending the adoption of memory safe languages to prevent such vulnerabilities [12]. While Rust is far from the only memory safe language in existence, it was selected for its emphasis on both memory safety and performance. When benchmarked against 155 other web frameworks, Rust's axum framework (which was selected for this project) was ranked 10th, performing at 79.1% relative to the most performant framework [13].

The game was built as a web application in order to maximize portability. This type of application can be run on any modern device with an internet connection, which is important for availability in resource-constrained environments. In contrast to a native application that runs on a device directly, outside of a web browser, web applications require no further work to be ported to different devices and operating systems. Furthermore, no additional installation process is required before the game can be played, increasing ease of adoption.

The database schema for the game consists of three tables: `games`, `challenges`, and `her2_cores`. A record in the `games` table consists of a unique, monotonically increasing integer identifier, a `username` field to store the player's name, timestamps for when the game was started and finished, the game's total score, the game's maximum possible score, and the total amount of time taken to complete all challenges. A record in the `challenges` table consists of a unique, monotonically increasing integer identifier, the identifier of the game to which the challenge belongs, the identifier of the challenge's associated HER2 core, the player's guess, timestamps for when the challenge was started and when its guess was submitted, and the number of points awarded for the challenge. A record in the `her2_cores` table consists of a unique, monotonically inreasing integer identifier, the HER2 score of the core patch, the name of the file storing the image of the patch, and a timestamp for when the patch was imported into the database. A many-to-one relationship is formed between challenges and patches, and a one-to-many relationship is formed between games and challenges. While the `points` and `score` fields are computed from other fields already

stored in the database and are thus duplicated data, we retain them to reduce the complexity of certain database queries, in particular the query to compute a game's total score and the query to generate the leaderboard.

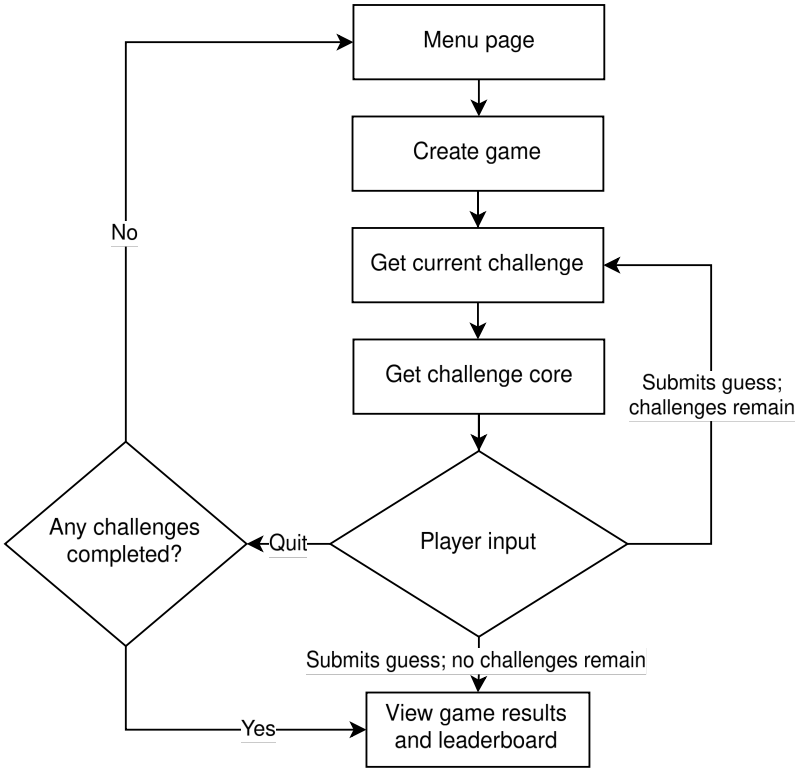

Figure 2: Flowchart of game progression

After the player enters their name on the menu page and clicks "Play," the game begins with an HTTP POST request to the `/games` endpoint on the backend application server. This creates a game record and 20 corresponding challenge records in the database. The player is then navigated to the game page, at which point a GET request is made to the `/games/:id/challenge` endpoint, where `:id` is replaced with the identifier of the game record. The endpoint returns the identifier of the next uncompleted challenge in the game (as well as the total number of challenges and completed challenges), and a subsequent GET request is made to `/challenges/:id/core` to get the image of the challenge's corresponding patch. The patch is then displayed to the player, and after a two second delay, the buttons for them to make their selection are enabled. The player can either indicate what they believe to be the HER2 status of the patch, or they can quit the game. If the player quits the game and no challenges have been completed, they are navigated back to the main menu page. If some challenges have been completed, the game's total score and duration are computed and stored, and the player is sent to the results page. When the player clicks one of the buttons to indicate the patch's HER2 score, a POST request is made to the `/challenges/:id` endpoint, which records the player's guess. Then, if any challenges remain (as indicated by the response from the initial GET `/games/:id/challenge` request), a new GET request is made to `/games/:id/challenge`, and the game continues. If all challenges have been completed, the user is sent to the results page. On the results page, a GET request is made to `/games/:id` to get the player's score, as well as each submitted challenge and its correct HER2 score. Then, a GET request is made to `/leaderboard` to fetch the leaderboard this process is illustrated in Fig. 2.

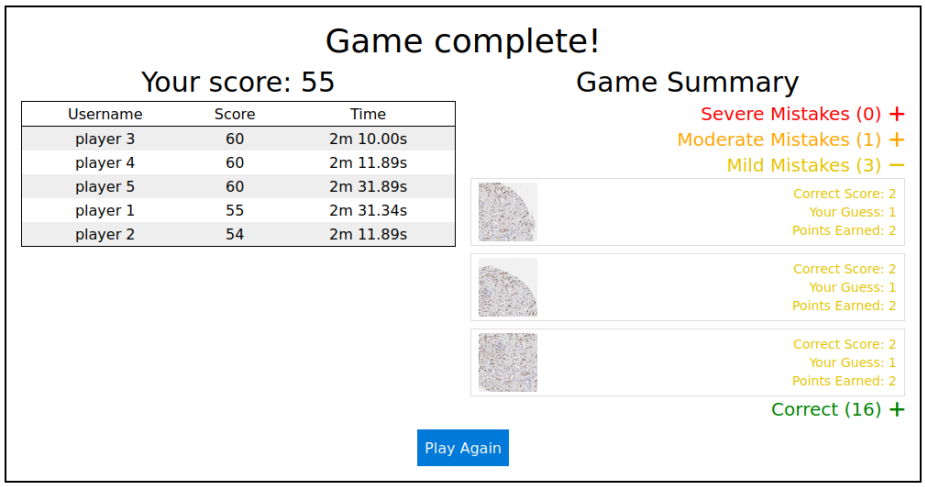

Figure 3: Results page

3.3.1 Design Decisions

Rather than allow the user to query a patch by its identifier directly, patches must be obtained through their associated challenges. This allows for greater control over when a patch is sent to the user, as once a challenge has been submitted, the backend server can reject any further requests for that challenge's patch. This is to obstruct malicious users who seek to download the entire body of HER2 data from the game. If HER2 patch identifiers were exposed, a malicious user could request one patch after another and eventually download the entire dataset. While this could be mitigated by encoding the identifiers in a non-sequential way, it would still leave open the option for users to access cores independently of a running game, which is still not intended behavior. While it is not possible to completely prevent users from amassing their own repository of patch images, this measure increases the difficulty of such behavior.

4 Discussion

Future work includes presenting the platform to a large sample of pathologists to determine its efficacy at developing diagnostic skills that can be applied to real-world cases. One option is a controlled study, in which two groups of pathologists of similar experience are selected, one partakes in a training session on the platform, and then both are evaluated on diagnostic accuracy of one or more real-world samples. Then, a significant improvement in accuracy in the experimental group would suggest that the platform is an effective training tool.

Furthermore, the BioGames platform need not be exclusive to the evaluation of HER2 IHC stained tissue samples. Any diagnostic technique in which a quantitative score is assigned to an image or set of images of a tissue sample could be trained using the platform. While the current dataset consists only of HER2 stained breast tissue, HER2 has also been approved as a diagonstic stain for gastric, esophageal, and lung cancers. A futher example of a candidate diagnostic criterion suitable for BioGames is PD-L1 over-expression. As IHC is a widely available, practical, and economical approach for studying PD-L1 expression in a tumor [14], and interpretations are given as a numeric tumor proportion score (TPS) [14], BioGames could be easily extended to offer training for PD-L1 diagnosis.

The challenge scoring function is one possible area of improvement. The absolute difference method described in the Results section is adequate for scoring challenges of HER2 classification. However, with the expansion of the platform to the diagnosis of other biomarkers, the need may arise for a standardized score that can be compared across different diagnostic techniques. In the case of PD-L1, assuming the absolute difference method were applied to the evaluated and true TPS's, scores would range from 0 to 100, while they would range from only 0 to 3 for

HER2. An alternative proposed scoring function is

$$\text{Score}(C, T, \Delta, t) = \frac{C\Delta}{1 + e^{t/T}}$$

where Δ is the absolute score difference, t is the amount of time taken for the challenge, and C and T are arbitrary scale factors. With manipulation of the scale factors for different biomarkers and scoring paradigms, this function can be altered to produce scores on an arbitrary interval. As such, the score difference and time taken can be appropriately weighted to generate scores comparable across different biomarkers and diagnostic techniques. Furthermore, both the player's accuracy and speed will be factored into a single numeric score. However, this may or may not be desirable, as one may want to place highly accurate players in a superior position to inaccurate players, regardless of speed.

In addition, the game has yet to be deployed to publicly accessible infrastructure. An intended starting point is a server on UCLA's SEASnet, but the growth of traffic and storage needs may warrant consideration of cloud computing platforms such as Amazon Web Services, Microsoft Azure, Google Cloud Platform, or Linode.

Finally, much work remains to bring BioGames up to the standards of a modern web application. One critical feature that remains to be implemented is proper authentication and authorization. This feature essential for the integrity of a user's scores, as users should not be able to impersonate each other. Additionally, proper user management will include the ability for users to sign in through their academic institution. This will enrich the competitive experience of the platform, as rankings can be held by both individual users and their institutions.

This version of BioGames represents a first step in the direction of widely available and easy-to-use digital pathology training platforms. With future work, it has the potential to improve the diagnostic skills of pathologists around the world, and decrease the amount of time necessary to reach an acceptable level of diagnostic acuity.

5 References

1. Mitri, Z., Constantine, T., & O'Regan, R. (2012). The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy. Chemotherapy Research and Practice. https://doi.org/10.1155/2012/743193

2. Raghav, K. P. S., & Moasser, M. M. (2023). Molecular Pathways and Mechanisms of HER2 in Cancer Therapy. Clinical Cancer Research. https://doi.org/10.1158/1078-0432.CCR-22-0283

3. Gutierrez, C., & Schiff, R. (2011). HER2: Biology, Detection, and Clinical Implications. Archives of Pathology & Laboratory Medicine. https://doi.org/10.5858/2010-0454-RAR.1

4. Duraiyan, J., Govindarajan, R., Kaliyappan, K., & Palanisamy, M. (2012). Applications of immunohistochemistry. Journal of Pharmacy & BioAllied Sciences. https://doi.org/10.4103/0975-7406.100281

5. Engelberg, J. A., Retallack, H., Balassanian, R., Dowsett, M., Zabaglo, L., Ram, A. A., Apple, S. K., Bishop, J. W., Borowsky, A. D., Carpenter, P. M., Chen, Y., Datnow, B., Elson, S., Hasteh, F., Lin, F., Moatamed, N. A., Zhang, Y., & Cardiff, R. D. (2015). "Score the Core" Web-based pathologist training tool improves the accuracy of breast cancer IHC4 scoring. Human Pathology, 46(11), 1694–1704. https://doi.org/10.1016/j.humpath.2015.07.008

6. Selcuk, S. Y., Yang, X., Bai, B., Zhang, Y., Li, Y., Aydin, M., Unal, A. F., Gomatam, A., Guo, Z., Angus, D. M., Kolodney, G., Atlan, K., Haran, T. K., Pillar, N., & Ozcan, A. (2024). Automated HER2 scoring in breast cancer images using deep learning and pyramid sampling. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2404.00837

7. facebook. (n.d.). Release v0.3.0 · facebook/react. GitHub. https://github.com/facebook/react/releases/tag/v0.3.0

8. TypeScript. (n.d.). CodePlex. https://web.archive.org/web/20150403224440/https://typescript.codeplex.com/releases/view/95554

9. Stonebraker, M., & Rowe, L. A. (1986). The design of POSTGRES. SIGMOD Record, 15(2), 340–355. https://doi.org/10.1145/16856.16888

10. Bugden, W., & Alahmar, A. (2022). Rust: the programming language for safety and performance. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2206.05503

11. National Security Agency. (2023). Cybersecurity Information Sheet - Software Memory Safety. Retrieved June 2, 2024, from https://media.defense.gov/2022/Nov/10/2003112742/-1/-1/0/CSI_SOFTWARE_MEMORY_SAFETY.PDF

12. The White House. (2024). Press Release: Future Software Should Be Memory Safe. Retrieved June 2, 2024, from https://www.whitehouse.gov/oncd/briefing-room/2024/02/26/press-release-technical-report/

13. TechEmpower. (n.d.). Web Framework Benchmarks. Retrieved June 2, 2024, from https://www.techempower.com/benchmarks/\#hw=ph\&test=compositeion=data-r22

14. Akhtar, M., Rashid, S., & Al-Bozom, I. A. (2021). PD-L1 immunostaining: what pathologists need to know. Diagnostic Pathology, 16(1). https://doi.org/10.1186/s13000-021-01151-x