Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction

Isabel Straw, Honghan Wu

ABSTRACT

Objectives The Indian Liver Patient Dataset (ILPD) is used extensively to create algorithms that predict liver disease. Given the existing research describing demographic inequities in liver disease diagnosis and management, these algorithms require scrutiny for potential biases. We address this overlooked issue by investigating ILPD models for sex bias.

Methods Following our literature review of ILPD papers, the models reported in existing studies are recreated and then interrogated for bias. We define four experiments, training on sex-unbalanced/balanced data, with and without feature selection. We build random forests (RFs), support vector machines (SVMs), Gaussian Naïve Bayes and logistic regression (LR) classifiers, running experiments 100 times, reporting average results with SD.

Results We reproduce published models achieving accuracies of >70% (LR 71.31% (2.37 SD) – SVM 79.40% (2.50 SD)) and demonstrate a previously unobserved performance disparity. Across all classifiers females suffer from a higher false negative rate (FNR). Presently, RF and LR classifiers are reported as the most effective models, yet in our experiments they demonstrate the greatest FNR disparity (RF: –21.02%; LR: –24.07%).

Discussion We demonstrate a sex disparity that exists in published ILPD classifiers. In practice, the higher FNR for females would manifest as increased rates of missed diagnosis for female patients and a consequent lack of appropriate care. Our study demonstrates that evaluating biases in the initial stages of machine learning can provide insights into inequalities in current clinical practice, reveal pathophysiological differences between the male and females, and can mitigate the digitisation of inequalities into algorithmic systems.

Conclusion Our findings are important to medical data scientists, clinicians and policy-makers involved in the implementation of medical artificial intelligence systems. An awareness of the potential biases of these systems is essential in preventing the digital exacerbation of healthcare inequalities.

BACKGROUND

Liver cirrhosis accounts for 1.8% of deaths in Europe, a number which has grown significantly over the past decade as rates of alcohol consumption, chronic hepatitis infections and obesity-related liver disease have increased. Yet, liver disease does not affect all populations equally. Recent research has demonstrated sex differences in the prevalence, diagnosis and management of various hepatic illnesses. A key determinant of patient outcomes from liver disease is the early detection of pathology, yet when it comes to diagnosis and referral, female patients appear to be at a significant disadvantage. In alcohol related liver disease, Vatsalya et al report that women are less likely to be
suspected of alcohol abuse, diagnosed and often experience more severe disease with worse outcomes. Sex differences in diagnosis are compounded by inequalities in the liver disease management. Mathur et al report disparities in access to liver transplantation that result in females having markedly lower transplant rates than their male counterparts. The problem extends beyond hepatology. In 2021, the UK parliamentary report on the gender health gap highlighted that the UK has the largest female health gap in the G20 and the 12th largest globally. The exclusion of females from research trials (extending to animal research), the neglect of female bodies throughout medical pedagogy and the unconscious biases of practitioners are a few of the intersecting factors that result in worse health outcomes for female patients.

Liver function tests are integral to patient diagnosis and monitoring. These ‘biochemical markers’ include proteins made by the liver (eg, albumin), and enzymes required for metabolism (eg, aspartate aminotransferase (AST)). Bias research has illustrated that biochemical markers are not equally effective for all patient groups. Suthahar et al describe how sex differences in biomarker thresholds affect objectivity in management, as what is considered ‘normal’ in one sex, may not be so in the other. Grimm et al investigate the relationship between albumin and mortality, reporting that albumin offers a higher predictive power for males compared with females. Furthermore, Vatsalya et al and Stepien et al describe sex differences in biochemical cut offs, highlighting that the milder expression of liver injury for females may result in female disease going undetected. Such disparities in the predictive potential of clinical biomarkers have the potential to exacerbate healthcare inequalities.

The rise in healthcare artificial intelligence (AI) has resulted in the increasing use of large clinical datasets for machine learning (ML). ML classifiers that use biochemical markers to model patient trajectories have consistently outperformed traditional statistical models. However, despite the promise of ML tools, the presence of demographic biases in AI algorithms has indicated that historical harms may materialise in digital systems and worsen population inequalities. The development of predictive models from biomarkers is one area in which medical ML models are at risk of encoding the errors of current practice. In our paper we explore for this possibility in liver disease prediction by examining models built from a commonly cited dataset: The Indian Liver Patient Dataset (ILPD).

The ILPD is a widely used open-source dataset that provides the biochemical markers of a sample of patients, some of whom have liver disease. BanuPriya and Tamilselvi provide an overview of classification models built from this dataset, since which time further models have been published from both academics and major industry. Authors consistently report accuracies of >70% for identifying liver patients, with logistic regression (LR) models and random forests (RFs) giving the best results. Jin et al demonstrate accuracies of 72.7% with LR models, similarly Adil et al achieve 74% accuracy with their LR model, outperforming artificial neural networks and support vector machines (SVMs). A recent study from Intel reproduces these models and performs additional feature selection giving model accuracies of 74.6% (RF) and 71.2% (SVM).

Predictive ML models may benefit patient care if they can diagnose liver disease at an earlier stage. Yet, despite the existing literature that describes biases in clinical medicine, biochemical tests and algorithmic performance, none of the ML studies on the ILPD focus on sex disparities in model performance. We seek to address this gap in the research by investigating the ILPD dataset and its respective models for sex bias.

### METHODOLOGY

The ILPD was originally collected from India and consists of 583 patient records, of which 416 have liver disease. We imported the ILPD from the UCI repository (full code available in online supplemental material C). Data exploration and initial analysis

Data exploration is the primary stage of the ML process and involves file importation, formatting, descriptive statistics and configuring datatypes. Online supplemental table 1 gives the variables included in our dataset and their initial datatypes.

Feature exploration

Online supplemental table 2 presents the sex-stratified feature importance ranked by Pearson’s correlation coefficient. For females, the enzymes ALT and AST are ranked fourth and fifth, whereas for males they are ranked seventh and eighth. Further, albumin and A/G ratio are ranked higher for male patients compared with female patients. These subtle differences in feature importance may reflect underlying sex differences in hepatic physiology and biomarker expression. Further, online supplemental table 2 demonstrates that the mean IQR across all biomarkers is less for females, suggesting that these biomarkers may have less of a predictive power for female patients overall (mean IQR; female 0.145, male 0.175).

Data preprocessing

Data preparation steps reflected existing studies. Mean imputation was used to address missing values, gender was mapped to a 0/1 numerical datatype, normalisation was performed using minimum-maximum scaler function and the target variable was recoded to binary variable, such that 1 represents diseased patients (n=416).

Addressing class imbalance

The original dataset demonstrated significant class imbalance (167 healthy vs 416 diseased patients) and sex imbalance (142 females vs 441 males). Similarly to existing models, we implement the imblearn SMOTE(1)}
package to address these imbalances; oversampling both the minority class and under-represented females as detailed in Table 1.19 The sex-unbalanced dataset is retained to compare the impact of female representation in the training data on sex disparities in performance.

**Model development and implementation**

Gulia and Praveen Rani review the classification algorithms that have been built from the ILPD, including RFs and SVMs.20 A more recent review from BanuPriya and TamilSelvi describe the accuracies of additional models including Bayesian Networks, which is further built on by the work of Aswathy who evaluates the performance of LR models on the ILPD.18 19 We replicate the methods of these studies, reproducing RF, SVM, Gaussian Naïve Bayes (GNB) and LR classifiers. We implement these models across four experiments, in which we evaluate the overall and sex-stratified performance of the classifiers.

**Experiment 1: models trained on unbalanced dataset, without feature selection**

Initially, we reproduce existing studies, building a predictive algorithm on the full unbalanced dataset to predict liver disease. Data were divided into test and training subsets (30%/70%), hyperparameters were tuned using GridSearchCV, the model was trained on the mixed-sex data and results were stratified by sex to give the evaluation metrics for males/females separately. We do this 100 times (building, training and testing separate models) and report average results with SD over the 100 runs. This is done for all four classifiers resulting in four results tables (online supplemental material B Spreadsheets, ‘Experiment 3.1.1—RF’—‘Experiment 3.1.1 GNB’).

**Experiment 2: models trained on sex-balanced dataset, without feature selection**

The methodology of experiment 1 is repeated using the sex-balanced dataset defined in Table 1. We ensure sex balance in the training data by taking random subsets from the male and females separately, which are appended together to form the full sex-balanced training data for each individual experiment (online supplemental file 3 Spreadsheets, ‘Experiment 3.1.2—RF’—‘Experiment 3.1.2 GNB’).

**Experiment 3: models trained on unbalanced dataset, with feature selection**

In experiment 3, we perform feature selection based on the unbalanced dataset, in experiment 4, we perform feature selection on the sex-balanced dataset. Feature selection is performed using Recursive Feature Elimination (RFE) package, which returns the top five ranked features (online supplemental material B Spreadsheets, ‘Experiment 3.1.3—RF’—‘Experiment 3.1.3 GNB’).

**Experiment 4: models trained on balanced dataset, with feature selection**

Lastly, models and feature selection are fitted to the sex-balanced dataset. Our aim was to investigate whether feature selection would differ once the representation of females was addressed, and whether this would influence any performance disparities.

**Model evaluation**

Evaluation metrics are reported for all patients and separately for the sexes (equations 1–3). We examine the mean difference between the male and females for each evaluation metric to demonstrate any disparities (equation 4). Two-sample paired t-tests are run on the series of 100 experiments for the male and female patients to assess whether the mean difference between sexes, for each of the evaluation metrics, is statistically significant (p<0.05).

**Equation 1: accuracy evaluation metric**

Accuracy gives the proportion of correct predictions produced by a model.

\[
\text{Accuracy} = \frac{\text{True positives} + \text{True Negatives}}{\text{True positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}
\]

**Equation 2: F-score evaluation metric, precision and recall**

The F-score is the average of precision and recall, with a value of 1 being a perfect score.

\[
\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}
\]

\[
\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}
\]

\[
\text{F Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

---

**Table 1** Summary counts of classes in the Indian liver patient dataset, including counts after the dataset is balanced

<table>
<thead>
<tr>
<th>Target (disease=1)</th>
<th>Dataset 1 (original)</th>
<th>Total counts for sexes</th>
<th>Dataset 2 (oversampled minority class)</th>
<th>Total counts for sexes</th>
<th>Dataset 3 (sex balanced, oversampled females)</th>
<th>Total counts for sexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0</td>
<td>50</td>
<td>142</td>
<td>145</td>
<td>237</td>
<td>408</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>92</td>
<td>92</td>
<td>92</td>
<td>187</td>
<td>187</td>
</tr>
<tr>
<td>Male</td>
<td>0</td>
<td>117</td>
<td>441</td>
<td>271</td>
<td>595</td>
<td>271</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>324</td>
<td>324</td>
<td>324</td>
<td>595</td>
<td>595</td>
</tr>
<tr>
<td>Total</td>
<td>583</td>
<td>832</td>
<td>1190</td>
<td></td>
<td></td>
<td>1190</td>
</tr>
</tbody>
</table>

---

Copyright © 2022 BMJ Publishing Group Ltd. All rights reserved. BMJ Health Care Inform: first published as 10.1136/bmjhci-2021-100457 on 25 April 2022. Downloaded from http://informatics.bmj.com/ on October 25, 2022 by guest. Protected by copyright.
Equation 3: performance error rates
The following error rates are used throughout our evaluation.21
- True positive: Predicted yes and they do have disease.
- True negative: Predicted no and they do not have disease.
- False positive: Predicted yes, but they do not have disease.
- False negative: Predicted no, but they actually do have disease.

\[
\text{True Negative Rate (TNR)} = \frac{TN}{TN+FP}
\]
\[
\text{True Positive Rate (TPR)} = \frac{TP}{TP+FN}
\]

Equation 4: sex performance disparity
Sex performance disparity = Male evaluation metric (mean) − Female evaluation metric (mean)

RESULTS
We ran 16 experiments: experiments 1–4, with each of the four classifiers. The detailed results tables with the 100 experiment runs are provided in the spreadsheet files in online supplemental material B. In online supplemental material A ‘Tables in Text’, we provide summary in several condensed tables, which give the average evaluation metrics and the statistical significance of any male-female differences.

Results for experiment 1
Online supplemental table 3 demonstrates that our four models reflect the existing literature, achieving accuracies >70% (71.31% (2.37 SD) LR – 79.40% (2.50 SD) SVM). Table 2 details the disparities for each evaluation metric, from which we observe a statistically significant sex disparity in Accuracy for all classifiers, with mixed results regarding the direction of the disparity (performance disparity −2.98 SVM to 2.96% RF, p<0.05). In the case of the ROC_AUC score, we observe a significant disparity that negatively impacts females for the RF (6.80%, p<0.05), LR (2.93%, p<0.05) and GNB (5.53%, p<0.05) classifiers.

The accuracy and ROC_AUC disparities fluctuate depending on the balance between the different error rates, however, on examining the error rates individually, we see a consistency in error trends for each sex. Across all classifiers females suffer from a higher false negative rate (FNR), while males suffer from a higher false positive rate. The disparity demonstrates a consistently higher recall for males, with females experience a lower recall and correspondingly higher FNR disparity, −2.58% to −24.07%, (table 2).

Results for experiment 2
In experiment 2, we trained on sex-balanced data, improving overall accuracy across all four classifiers (RF 81.66% (2.33 SD) vs 78.17 (2.36 SD); LR 74.53% (1.96 SD) vs 71.31% (2.37 SD); SVM 83.30% (1.75 SD) vs 79.40% (2.50 SD); GNB 74.75% (1.9 SD) vs 71.53% (2.61 SD)—online supplemental table 4). We now see a consistent accuracy disparity that benefits females across all four classifiers (−11.47% to −6.17%, p<0.05—table 3). Disparities in the ROC_AUC scores are less consistent (LR unbalanced ROC disparity 2.93%, LR balanced ROC disparity 4.79%; GNB unbalanced ROC disparity 5.53%, GNB balanced disparity 5.45%).

Table 2
Experiment 3.1.1—unbalanced training data without feature selection, sex performance disparities

<table>
<thead>
<tr>
<th>Mean difference averaged over n=100</th>
<th>Random forest classifier</th>
<th>Logistic regression classifier</th>
<th>Support vector machine</th>
<th>Gaussian Naïve Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sex performance disparities (%)</td>
<td>t-test p value</td>
<td>Sex performance disparities (%)</td>
<td>t-test p value</td>
</tr>
<tr>
<td>Accuracy</td>
<td>2.96</td>
<td>0.00</td>
<td>−2.85</td>
<td>0.01</td>
</tr>
<tr>
<td>FScore</td>
<td>15.63</td>
<td>0.00</td>
<td>15.86</td>
<td>0.00</td>
</tr>
<tr>
<td>ROC_AUC*</td>
<td>6.80</td>
<td>0.00</td>
<td>2.93</td>
<td>0.00</td>
</tr>
<tr>
<td>Precision</td>
<td>5.25</td>
<td>0.00</td>
<td>−4.87</td>
<td>0.00</td>
</tr>
<tr>
<td>Recall</td>
<td>21.02</td>
<td>0.00</td>
<td>24.07</td>
<td>0.00</td>
</tr>
<tr>
<td>False negative rate</td>
<td>−21.02</td>
<td>0.00</td>
<td>−24.07</td>
<td>0.00</td>
</tr>
<tr>
<td>True negative rate</td>
<td>−7.42</td>
<td>0.00</td>
<td>−18.20</td>
<td>0.00</td>
</tr>
<tr>
<td>False positive rate</td>
<td>7.42</td>
<td>0.00</td>
<td>18.20</td>
<td>0.00</td>
</tr>
<tr>
<td>True positive rate</td>
<td>21.02</td>
<td>0.00</td>
<td>24.07</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*ROC AUC score is a measure of the separation between classes in a binary classifier, derived from the area under the ROC curve.
Online supplemental table 5 presents a comparison of the evaluation metrics with/without balancing of the training data. In one case, we observe an improvement in performance for all patients. When trained on the balanced dataset, the LR accuracy improves overall (74.53% (1.96 SD) vs 71.31% (2.37 SD)), for females (77.71% (2.42 SD) vs 73.33% (3.95 SD)) and for males (71.35% (3.22 SD) vs 70.49% (2.74 SD)).

**Results for experiment 3**

We did not see an improvement in overall performance or a reduction in disparities with RFE. A significant ROC_AUC disparity is apparent across all four classifiers (3.60%–6.61%, p<0.05) that negatively impacts females. We see the same error rate findings as earlier, with a higher FNR for females (FNR Disparity −18.21 to −21.24%, p<0.05, table 4 and online supplemental table 6).

**Results for experiment 4**

Experiment 4 gives mixed results. The accuracy disparity benefits females across all classifiers (−4.64% to −6.80%, p<0.05), whereas the ROC_AUC disparity demonstrates a benefit for males in three out of four classifiers (−0.05% to 5.95%, p<0.05, table 5) The results relate to the subtle changes in error rates with each model, however, across
all classifiers the FNR is consistently higher for females (−9.70% to −22.78%, p<0.05 (online supplemental table 7)).

Analysis of feature selection

Online supplemental table 8 gives the feature rankings assigned by the RFE model when fitted to unbalanced and balanced data, focusing on RF classifiers. When we address the under-representation of females in the training data, ALP and gender are included as the top two features, while A/G ratio and total bilirubin are removed. This finding may reflect existing research that describes sex differences in biomarker expression. In their analysis gender-specific intervals for hepatic biomarkers, Li et al highlight sex differences in ALP, ALT and GGT, indicating that differing thresholds may be appropriate for diagnosis. Sex differences in biochemical disease profiles may explain why integrating more female patients affects the feature selection in experiment 4.

Table 5: Experiment 3.1.4—balanced training data with feature selection, sex performance disparities

<table>
<thead>
<tr>
<th>Feature</th>
<th>Random forest classifier</th>
<th>Logistic regression classifier</th>
<th>Support vector machine</th>
<th>Gaussian Naive Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sex performance disparities (%)</td>
<td>t-test p value</td>
<td>Sex performance disparities (%)</td>
<td>t-test p value</td>
</tr>
<tr>
<td>Accuracy</td>
<td>−5.62</td>
<td>0.00</td>
<td>−6.80</td>
<td>0.00</td>
</tr>
<tr>
<td>FScore</td>
<td>7.86</td>
<td>0.00</td>
<td>14.39</td>
<td>0.00</td>
</tr>
<tr>
<td>ROC_AUC</td>
<td>−0.05%</td>
<td>0.46</td>
<td>3.57%</td>
<td>0.00</td>
</tr>
<tr>
<td>Precision</td>
<td>4.60%</td>
<td>0.00</td>
<td>9.28%</td>
<td>0.00</td>
</tr>
<tr>
<td>Recall</td>
<td>9.70%</td>
<td>0.00</td>
<td>15.51%</td>
<td>0.00</td>
</tr>
<tr>
<td>False negative rate</td>
<td>−9.70</td>
<td>0.00</td>
<td>−15.51</td>
<td>0.00</td>
</tr>
<tr>
<td>True negative rate</td>
<td>−9.79</td>
<td>0.00</td>
<td>−8.37</td>
<td>0.00</td>
</tr>
<tr>
<td>False positive rate</td>
<td>9.79</td>
<td>0.00</td>
<td>8.37</td>
<td>0.00</td>
</tr>
<tr>
<td>True positive rate</td>
<td>9.70</td>
<td>0.00</td>
<td>15.51</td>
<td>0.00</td>
</tr>
</tbody>
</table>

DISCUSSION

In recent years, research has highlighted that medical biases and female under-representation may significantly contribute to differences in healthcare outcomes; in our paper, we have examined how this phenomena may extend into ML. We present several key findings:

► Model reproduction and demonstration of disparity:
  We have demonstrated a previously unobserved sex disparity that exists in published ML classifiers based on the ILPD dataset.

► Error disparities: Sex disparities in Accuracy and ROC_AUC fluctuate depending on model and the balance between error rates, however, sex differences in specific error rates are consistent. We observe a consistently lower recall and correspondingly higher FNR for females. Of note, RF and LR classifiers are reported as the most effective on the ILPD dataset, however, these models demonstrate the greatest disparity in the FNR when trained on the original dataset (RF, FNR disparity −21.02% (p<0.05); LR, FNR disparity −24.07%, (p<0.05)). Clinically, this FNR disparity would materialise as an inequality in disease detection that negatively impacts females, with higher instances of missed disease.

► Balanced training: Training on sex-balanced data improved overall performance for all classifiers. In the case of the LR classifier, accuracy improves overall and for the sexes separately, indicating that with the right model selection addressing poor performance for the under-represented group does not need to come at the expense of the majority group.

► Impact of model architecture on disparity: Our experimental outcomes were not consistent across models, indicating that bias mitigation techniques may need to be tailored to model choice.

► Analysis of feature ranking: Our comparison of feature importance reinforces existing clinical research that highlights the sex differences in the role of liver biomarkers.

Implications for data science

Our experiments demonstrated that sex-specific feature selection and addressing under-representation of females may be an important bias mitigation technique when developing ML algorithms in medicine. Furthermore, we illustrate that there is no consistent solution across all classifiers, suggesting techniques need to be tailored
to model choice. ML models also present novel opportunities for improving existing practice and addressing health disparities that relate to biochemical discrepancies between the sexes. Given the evolving evidence that critiques the use of ‘unisex’ biochemical thresholds, ML models that do not rely on these defined thresholds may pose a superior alternative if developed with an awareness of the subtle sex differences in disease manifestation.

Implications for clinical medicine and public health
Classification algorithms are being increasingly used in healthcare settings to assist clinicians in medical diagnosis. Unless these algorithms are evaluated for biases, they may only improve care for a subset of patients and consequently increase healthcare inequalities. By evaluating ML models for demographic biases before they are implemented in digital medicine, we can mitigate the perpetuation of these inequalities into digital systems.

Furthermore, insights from model development can be used to inform current clinical care. Our data exploration of feature correlation demonstrated sex differences in feature importance. Such research can inform practising clinicians on the relevance of different indicators for the patient in front of them, for example, albumin may be more indicative of pathology in males. Lastly, examining disparities in algorithmic performance offers an opportunity to reflect on which patients may be being missed in current practice. Throughout our analysis, we demonstrated a persistently high FNR for females, suggesting that female disease is at risk of being overlooked. Examining the physiological profile of algorithmic false negatives presents an opportunity to better understand which patients are at risk of being misdiagnosed.

It should be noted that the ILPD does not include demographic information on race or ethnicity. Racial patients are at risk of being misdiagnosed. The historic absence of women from the healthcare system means that female disease is at risk of being overlooked. Examining the physiological profile of algorithmic false negatives presents an opportunity to better understand which patients are at risk of being misdiagnosed.

REFERENCES


M. BanuPriya, Tamilselvi PR. Performance analysis of liver disease prediction using machine learning algorithms. 5, 2018.


Auxilla LA. Accuracy prediction using machine learning techniques for Indian patient liver disease. 2nd International Conference on Trends in Electronics and Informatics (ICOEII), 2018: 45–50.


