

The accuracy trap

When impressive machine-learning results conceal consequential failures

David J. Ranney

Orcid iD: 0009-0000-4572-836X

Global Health Executive Mentoring (GHEM)

SMEC | Applied Research & Healthcare Leadership

October 2026

What if an AI model achieves 85% accuracy but fails to identify most of the cases it was designed to detect?

Interesting, but what are we talking about?

Imagine this. A data scientist builds a model to predict whether employees will stay with an organization. A straightforward yes-or-no question. Of course, there are many variables that might influence the outcome: management, salary, benefits, working conditions, and so forth. But for now, let's focus on the question: **Will an employee stay or leave?**

Our data scientist presents the model to the executive team. It achieves approximately 85% accuracy.

Sounds promising. So what?

Think about the potential impact. What if your HR team could reliably identify employees likely to leave? Perhaps the organization could improve retention, reduce recruitment and training costs, and address workplace problems before valuable employees depart.

Excellent! We're ready to push the *'go ahead'* button.

Right?

Well, not so quickly.

Let's rein in our excitement and ask what that 85% accuracy means. It's a respectable-looking number. Not perfect, certainly, but good enough to attract attention.

Before committing organizational resources, however, shouldn't we understand what the model is identifying?

Why? Doesn't 85% accuracy tell us the model is performing well?

What if I told you the following?

- Approximately 87% of employees in our test data stayed with the organization.
- Of the 39 employees who actually left, the model correctly identified just seven.
- A model that simply predicted *everyone will stay* would have achieved approximately 87% accuracy—better than our more sophisticated model.

Are you as excited to push that button now?

When the majority hides the problem

I can almost hear the executive team's response: *If simply predicting everyone will stay produces better accuracy, why are we paying for a sophisticated model?*

Fair question.

But the answer isn't necessarily to abandon the model. It's to understand what we're asking it to accomplish.

Our employee data contains considerably more people who stayed than people who left. This is called **class imbalance**: one outcome occurs much more frequently than the other.

Think of credit-card fraud. Most transactions are legitimate. A system that labels every transaction legitimate might appear highly accurate, but it would fail to detect the very activity we're concerned about.

Our employee model faces a similar challenge.

It correctly identifies many employees who stay. But it misses 32 of the 39 employees who leave. That's approximately 82% of the departures in our test data.

Now consider the alternative. The simplistic model that predicts everyone will stay achieves higher overall accuracy, but it identifies absolutely no departures.

So which model is better?

That depends on what decision we're trying to make.

If our goal is simply to maximize the percentage of correct predictions, the majority-only approach wins. But if we're trying to identify employees who may leave so we can intervene, neither accuracy figure is sufficient to determine whether the model is useful.

We need to understand what the model gets right, what it gets wrong, and whether its predictions support a worthwhile organizational response.

When the consequences become clinical

Let's move from employee retention to healthcare.

In a separate educational machine-learning exercise, we developed a model to classify COVID-19 cases as positive or negative.

The model achieved **86.83% accuracy**.

Once again, that sounds encouraging.

But let's look a little closer.

Of the 1,303 cases that were positive, the model correctly classified 966. The remaining 337 were incorrectly classified as negative.

In other words, the model missed approximately **26% of the positive cases**.

Now, before anyone becomes alarmed, this was a classroom exercise, not a clinically validated diagnostic system. These results illustrate how model evaluation works; they do not establish the performance of a real-world COVID screening tool.

But imagine a healthcare executive being presented with a comparable result from a proposed clinical system.

Would an overall accuracy of nearly 87% be enough to authorize its use?

Probably not without further investigation.

What happens when a positive case is classified as negative? What happens when a negative case is incorrectly flagged as positive? Do those errors carry the same clinical, operational, or financial consequences?

Of course not necessarily.

And that's precisely the point.

The importance of a prediction depends not only on whether it is correct, but also on what happens when it is wrong.

The executive's responsibility

Here's where healthcare leadership enters the discussion.

We don't need every hospital executive to become a data scientist. Nor should we expect executives to personally calculate every performance metric or understand every line of code.

But we should expect them to ask informed questions before accepting a model's recommendations or authorizing its use.

For example:

- **What are we trying to accomplish?** Is the model identifying the outcomes that matter to our patients or organization?

- **What happens when it gets something wrong?** Are the consequences acceptable, and what safeguards are in place?
- **What evidence supports its use?** Has the model been evaluated with the population, circumstances, and decisions for which we intend to use it?

These are not questions intended to undermine the work of data scientists. Quite the opposite. They connect technical expertise to the decisions that expertise is supposed to support.

A data scientist may demonstrate that a model performs as designed. That does not automatically establish that the model is appropriate for a particular clinical or operational decision.

That judgment requires collaboration between technical specialists, clinicians, and organizational leaders.

And leadership cannot simply hand off its own responsibility for the consequences.

Accuracy is a starting point, not a decision

Let's return to our original question.

What if a model achieves 85% accuracy but fails to identify most of the cases it was designed to detect?

We now know that both statements can be true.

Accuracy isn't meaningless. It's useful information. But it's only one part of understanding whether a model is suitable for its intended purpose.

The next time someone presents an impressive machine-learning result, resist the temptation to immediately push the '*go ahead*' button.

Instead, ask: **What does this number tell us, what doesn't it tell us, and what are the consequences if we get this decision wrong?**

Because technical expertise can be delegated.

Leadership accountability cannot.

Related articles:

When accuracy isn't enough: A machine-learning case in healthcare decision-making (SMEC, 2026)